Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/120233

Modelling the Misuse of Alcohol and Drugs in South Africa Using Bayesian Binary Logistic Regression

Makwelantle A. Sehlabana[1], *Daniel Maposa*[1,*] *and Alexander Boateng*[2]

[1]Department of Statistics and Operations Research, University of Limpopo, Polokwane, South Africa
[2]Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Received: 7 Apr. 2022, Revised: 11 Aug. 2022, Accepted: 3 Sep. 2022
Published online: 1 May 2023

Abstract: The misuse of alcohol and drugs is a continuous life threat globally, including in South Africa. For that reason, researchers continue to investigate the risk factors associated with alcohol and drugs misuse. Most studies in literature employed the classical logistic regression model to investigate these risk factors. However, some of the issues pertaining to the classical methods are accounted for in the Bayesian framework. Likewise, the Bayesian logistic regression model can also account for problematic issues to the classical logistic regression model. Several studies used the Bayesian logistic regression model to investigate the risk factors associated with alcohol and drugs misuse. Usually, most Bayesian studies utilize default prior probability distributions such as Jeffereys' prior and Zellner's informative g-prior distributions. Not long ago, modified versions of Zellner's informative g-prior distribution have been proposed. This study aims to evaluate the effectiveness of a modified Zellner's informative g-prior distribution and subdue separation in modelling the misuse of alcohol and drugs. The model developed through the use of a modified Zellner's g-prior distribution is compared to the models developed through the use of a hyper g-prior distribution and mixtures of g and n prior distribution. Comparisons are based on precision and average prediction error. Although the models yielded similar results, the modified version of Zellner's informative g-prior distribution resulted in narrow credible intervals, and a small everage prediction error. Separation is also accounted for in the model. In this study, the modified version of Zellner's informative g-prior distribution is evidently effective. All models are developed using the Bayesian adaptive sampling (BAS) R package. Further research may include evaluating some of the recommended prior distributions for the Generalised Linear Models (GLM) and comparison of Bayesian binary logistic regression developed in this study with logistic regression in Machine Learning algorithms.

Keywords: Bayesian framework; classical methods; logistic regression model; Misuse of alcohol and drugs; Zellner's g-prior

1 Introduction

The misuse of alcohol and drugs is a global health threat. McLellan [1] defined substance misuse as the use of alcohol, illegal drugs, and prescribed medication in ways that cause harm to ourselves, people around us and society. In 2016, alcohol misuse accounted for 3 million (5.6%) deaths worldwide [2]. The mortality rate due to alcohol misuse was higher than that caused by tuberculosis, HIV/AIDS, and diabetes [3]. In 2017, the misuse of alcohol and drugs accounted for 11.8 million deaths globally [4]. According to the United Nations Office on Drugs and Crime (UNODC), the number of people aged 16 to 64 who used drugs in 2016 was 30% higher when compared to the number of people of the same age group who used drugs in 2009 [5]. In South Africa, 80% of young male deaths related to alcohol and drugs consumption is estimated to be two times more than the world norm [6]. Cannabis is regarded as the primary drug among South Africans younger than 20 years old [7]. South African provinces dominant with misuse of drugs are Western Cape, Limpopo and Mpumalanga, whereas Eastern Cape, Free State, Northern Cape, and North West are dormant with alcohol misuse [7].

The tragedy with the misuse of alcohol and drugs is that the impacts are beyond the individual misuser. Society is included in harms such as drink and driving crashes, violence, crime, decreased work productivity, increased stress among health systems, damage to economies, and increased mortality [8]. According to [5], research suggests that the

* Corresponding author e-mail: daniel.maposa@ul.ac.za

adolescents (12-14) stage is a risk period for the commencement of drugs and alcohol use, while the drugs and alcohol use may peak among youths aged 18 to 25 years. Symptoms of substance misuse include repeated substance misuse related to legal problems, the use of a substance in a physically hazardous environment or situations, and repeated use despite having social and interpersonal problems caused by the effects of the substance misuse [9]. The medical issues caused by the misuse of drugs and alcohol addiction include cardiovascular disease, stroke, cancer, and mental disorders [10].

Risk factors associated with drugs and alcohol misuse include, amongst others: age, education level, employment status, race, marital status, peer pressure, and family background. The study of Branstrom and Andreasson [11], and the study of Mafa et al.[12] examined the drinking patterns among males and females through a survey report. Both studies revealed that males misuse alcohol than females. Mbandlwa and Dorasamy [13] investigated factors contributing to substance abuse in South Africa and identified unemployment, lack of family values and poor parenting guidance as risk factors that contribute to substance abuse. A similar study highlighted that the population group at high risk of alcohol disorder include adolescents, unmarried, and people with high income [14]. Several researchers used the classical logistic regression model to identify the risk factors associated with drugs and alcohol misuse. Vythilingum et al. [15] used Analysis of Variance and logistic regression to study the prevalence of alcohol and substance use in South Africa. Similary, the study in [14] also employed the logistic regression. However, Bayesian logistic regression model can account for problematic issues for the maximum likelihood estimation (MLE) method, used to develop classical models. The common problem in classical logistic regression is separation (i.e. the situation in which the ML estimate of at least one regression coefficient does not exist), which can be accounted for in the Bayesian methods [16]. Suleiman et al. [17] also highlighted that Bayesian approaches produce more stable parameters than the MLE method.

The present study is paramount because a Bayesian version of logistic regression will be developed to model the risk factors associated with the misuse of alcohol and drugs in all the provinces of South Africa. This model will use the g-prior, which Zellner first introduced in 1983. Hence it is commonly known as Zellner's g-prior. However, variants of g-priors for logistic regression models have been proposed lately by, for instance, Hanson et al. [18] proposed a version of a g-prior, on the basis that the prior distribution on the overall probability of success is set to match a beta distribution. Also, Lally [19] explored several methods for selecting hyper parameters for the g-prior distribution.

2 Materials and methods

This section presents the data source, study area and statistical techniques used to analyse the data.

2.1 Study frame and data collection

This study analyses the factors associated with the misuse of alcohol and drugs in South Africa. South Africa consists of nine provinces: Eastern Cape, Free State, Gauteng, Kwa Zulu-Natal, Limpopo, Mpumalanga, Northern Cape, North West, and Western Cape. The data used in this study is obtained from Statistics South Africa (StatsSA)(Accessible at: http://nesstar.gov.za:828/webview/). This data was extracted from StatsSA 2019 General Household Survey (G.H.S.). The survey's target population comprises all households in all nine South African provinces and residents of workers' hostels. The sample design for the 2019 G.H.S. used a two-stage sampling method. The first stage is the stratified design with probability-proportional-to-size (P.P.S.) sampling of primary sampling units (P.S.U.s) from each stratum and the second stage is the systematic sampling of dwelling units (D.U.s) from the sampled P.S.U.s. A sample self-weighting design was used at the provincial level. The extracted data consists of gender, alcohol and/or drugs misuse, provinces, population group, age, marital status and educational level.

2.2 The Logistic regression model, g-prior, and the determination of a posterior distribution

This section presents the theoretical transition from the classical logistic regression model to the introduction of a g-prior and the determination of a posterior distribution.

2.2.1 The logistic regression model

The data used in this study results from n independent and identical trials with two possible outcomes, success or failure. This implies that the response variable is binary. Suppose that θ denotes the probability of success and y denotes the

number of successes (the response "yes" is taken as success) in those n independent and identical trials. Then the probability distribution of y is given by:

$$P(y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, y = 0,1,...,n. \tag{1}$$

From (1), we are interested in the model for parameter θ, the logistic regression model. Suppose that $\mathbf{X_i}$ is the covariate vector of length p. Then, the logistic regression for the parameter θ in (1) is given by:

$$\theta_i(\mathbf{X_i}) = \frac{\exp(\mathbf{X_i'}\beta)}{1+\exp(\mathbf{X_i'}\beta)}, i = 1,2,...,n, \tag{2}$$

and the likelihood function of the logistic regression in (2) is given by:

$$L(y_i,\beta) = [\sum_{i=1}^{n}(y_i\mathbf{X_i'}\beta - h(\exp(\mathbf{X_i'}\beta)))], \tag{3}$$

where $\mathbf{h}(\mathbf{x}) = \log \mathbf{1} + \mathbf{e^x}$, for$x \in \mathbb{R}$.

2.2.2 Informative g-prior and the posterior distribution for logistic regression

The g-priors, also known as the maximal data information priors (MDIP), were introduced by Arnold Zellner in 1983. The g-priors are derived by maximising the average data information in the data density relative to the information in the prior density. However, they can also be interpreted as priors that maximise the logarithm expectation of the ratio of the likelihood function [20]. Suppose that $\mathbf{f}(\mathbf{x}|\theta)$ is a density of $\mathbf{x}$ given θ, and $\pi(\theta)$ is a prior density. Then, the joint density of $\mathbf{x}$ and θ is given by:

$$p(x,\theta) = f(x|\theta)\pi(\theta). \tag{4}$$

The negative entropy of $p(x,\theta)$ in (3) relative to a measure of information in $p(x,\theta)$ is denoted by:

$$-H = \iint p(x,\theta)\ln p(x,\theta)dxd\theta = \int [f(x|\theta)\ln f(x|\theta)dx]\,\pi(\theta)d\theta + \int \pi(\theta)d\theta = \int I(\theta)\pi(\theta)d\theta + \int \pi(\theta)\ln\pi(\theta)d\theta, \tag{5}$$

where $I(\theta) = \int (x|\theta)\ln f(x|\theta)dx$ is the negative entropy of $f(x|\theta)$. Therefore, the negative entropy in (5) is the average information in the data density plus the information in the prior density. This implies that the prior density must maximise the function given by:

$$G[\pi(\theta)] = \int I(\theta)\pi(\theta)d\theta - \int \pi(\theta)\ln\pi(\theta)d\theta. \tag{6}$$

Suppose that we write the joint density in (4) in the form:

$$p(x,\theta) = g(x|\theta)g(x).$$

Then $G[\pi(\theta|x)]$ can also be written as:

$$G[\pi(\theta)] = \int \left\{ \int \left[g(x|\theta)\ln\left(\frac{I(\theta|x)}{\pi(\theta)}\right)\right] d(\theta)\right\} h(x)d(x),$$

where $I(\theta|x)$ is the likelihood function. $G[\pi(\theta)$ in (5) can be maximised if $\int I(\theta)\pi(\theta)d(\theta) = 1$ [21]. This results in a solution:

$$\pi(\theta) \propto e^{I(\theta)}.$$

Let us consider the binomial distribution in (1) with a single Bernoulli trial. The observation is given by:

$$f(y,\theta) = \theta^y(1-\theta)^{1-y}.$$

Then, the negative entropy for this observation is given by:

$$I(\theta) = E(y)\ln\theta + E(1-y)\ln(1-\theta) = \theta\ln\theta + (1-\theta)\ln(1-\theta),$$

which results in a prior density $\pi(\theta) \propto e^{I(\theta)} = \theta^{\theta}(1-\theta)^{1-\theta}$.
Zellner's method is not parameterisation invariant [22]. Therefore, this study em-ploys a modified version of Zellner's g-prior. This modified version of Zellner's g-prior for logistic regression was proposed by [23] and later developed by [18]. For the logistic regression model in (2), the proposed prior density is denoted by:

$$\beta \propto N_p\left(0, 4n(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\right),$$

where β is a $1 \times p$ vector of coefficients, $\mathbf{X}$ is an $n \times p$ design matrix, $g = 4$ is a g-prior constant, and $\mu = 0$. This implies that the g-prior for logistic regression developed in this study is given by:

$$\pi(\beta) \propto e^{\frac{1}{2}\beta' v^{-1}\beta}, \tag{7}$$

where $v = 4n(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$.The posterior distribution through the Bayes theorem is defined as:

$$p(\beta, y) \propto \exp\left[\sum_{i=1}^{n}\left(y_i\mathbf{X}_\mathbf{i}'\beta - h(\exp(\mathbf{X}_\mathbf{i}'\beta))\right)\right] \times \exp\left(\frac{1}{2}\beta' v^{-1}\beta\right), \tag{8}$$

which is the product of the likelihood in (3) and the prior distribution (7). Using the posterior in (8) and Bayesian adaptive sampling (BAS) algorithm to sample the posterior distribution without replacement facilitates model estimations. This algorithm is implemented in the R package called BAS.

3 Results and Discussion

The results for Bayesian logistic regression model are presented and discussed in this section. Table 1 shows a description of all the variables, Table 2 presents the Generalised Variance Inflation Factors, Table 3 displays the credible intervals and the average prediction errors, and Table 4 presents the model used for inference in the study. The descriptive statistics are summarised through cross tables and graphs presented in the appendix.

3.1 Results

Table 1: Variable description.

Variable	Description	Data set code
Alcohol or drugs misuse	Binary response from the participants of a survey: yes or no	Alc_drug
Provinces	The nine provinces of South Africa: Eastern Cape (E.C.), Free State (F.S.), Gauteng (G.P.), Kwa-Zulu Natal (KZN),Limpopo (L.P.), Mpumalanga (M.P.), Northern Cape (N.C.), North West (N.W.), and Western Cape (W.C.).	Province
Gender	Gender of a participant: Female and male.	Gender
Population group	The ethnical background of a participant: African, Coloured, Indian/Asian, and White.	Race
Age	The participants' age group: Youth (12-24), Adults (25-64), and Elderly (65 years and over).	Age
Marital status	The marital status of the participants: Divorced, Married, and Single.	MaritlalS
Education level	The highest education level reached by the participants: Primary, Secondary, Tertiary, and other.	EduLevel

3.1.1 Assumptions of a binary logistic regression

A classical logistic regression model was developed and used to test whether the assumptions of using the logistic regression are violated or not.The response variable is binary as described in Table 1. There are no continuous variables in the data set. Hence there are no concerns regarding the assumption of linearity. Generalised Variance Inflation factors presented in Table 2 are used to test for multicollinearity between the independent variables. The GVIFs for each variable exceeds four. This indicates that the explanatory variables are not correlated. No influential observations were detected and each variable has got at least 10 cases of the response, this is presented through the cross tables and graphs in the appendix, and implies that the data is large enough. The assumptions of logistic regression model are not violated.

Table 2: Multicollinearity test.

Variable	GVIF
Province	1.7108
Gender	1.1701
Age	1.2629
Race	1.4642
Marital status	1.2829
Education level	1.247

3.1.2 Variable selection

The Least Absolute Shrinkage and Selection Operator (LASSO) method is used for variable selection. The results are presented on the Receiver Operating Characteristic (ROC) curves displayed in Figure 1. The figure shows that the Area Under the Curve (AUC) for a classical logistic regression model with all the explanatory variables is 0.8419 and the AUC for a classical logistic regression model with the insignificant variables removed is 0.7783. The model with the highest AUC is a good model compared to the model with the low AUC. For that reason, all explanatory variables will be used to develop a Bayesian logistic regression model in this study.

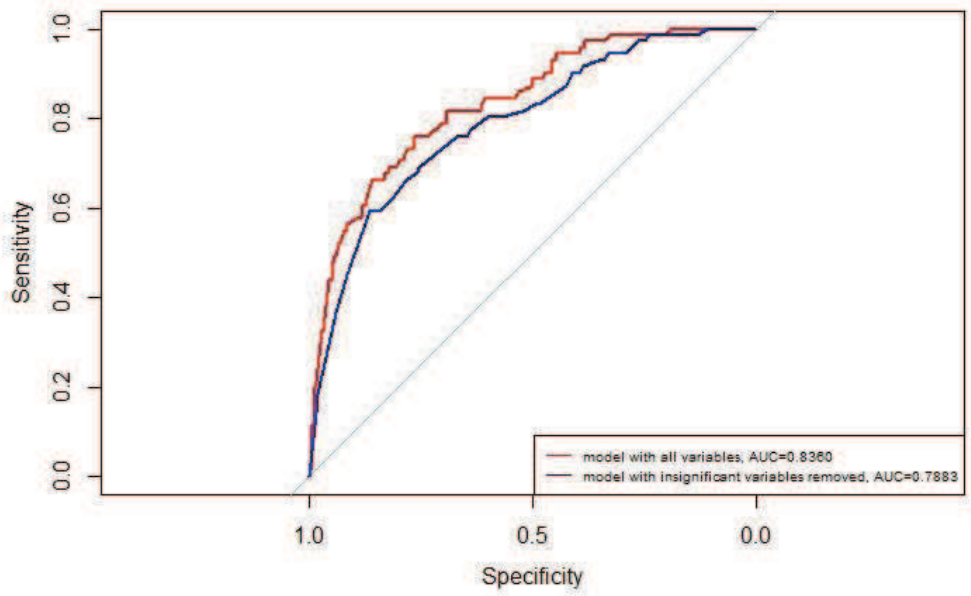

Fig. 1: ROC curves comparing two classical logistic regression models

3.1.3 Comparison of g-prior, hyper g-prior and mixtures of g/n prior distributions

The selected classical logistic regression model is separated. See the appendix for detection of separation test results. In this subsection, the association between the response variable and the explanatory variables is explored through the use of a modified Zellner's g prior distribution described in section 2 together with the hyper g-prior and mixtures of g/n priors recommended in [24]. The 95% credible intervals for the models developed through each prior distribution are presented in Table 3. The precision of credible intervals and the average prediction errors are used to compare the models. The average prediction errors are also used to test the models.

Table 3: The credible intervals(95%) for logistic regression models with hyper g-prior, mixtures of g/n prior and Zellner's g-prior.

Covariates	Hyper g-prior	Mixtures of g/n prior	Zellner' g-prior
Intercept	[-6.236: -6.346]***	[-7.617: -5.829]***	[-6.778: -5.383]***
Free State	[0.364: 1.646]**	[1.452: 1.633]***	[0.179: 1.354]**
Gauteng	[-1.463: 0.138]	[-1.554: 0.006]	[-1.637: -0.116]**
Kwa Zulu-Natal	[-0.454: 0.819]	[-1.211: 0.449]	[-0.649: 0.534]
Limpopo	[-1.102: 0.601]	[-3.384: 0.146]	[-1.357: 0.380]
Mpumalanga	[-0.483: 1.036]	[-1.961: 0.832]	[-0.660: 0.732]
Nothern Cape	[-0.649: -0.055]**	[-1.566: 0.401]	[-1.621: 0.279]
North West	[-0.714: 1.448]	[-1.297: 1.628]	[-0.479: 1.520]
Western Cape	[-3.554: -0.182]**	[-3.064: 0.000]	[-3.418: -0.216]**
Easter Cape(Ref)	- - - -	- - - -	- - - -
Male	[0.896: 2.103]**	[1.065: 2.309]***	[0.947: 2.197]**
Female(Ref)	- - - -	- - - -	- - - -
Elderly	[-1.344: 0.505]	[-3.186: 0.139]	[-1.366: 0.481]
Youth	[-2.254: -0.734]**	[-2.478: -0.894]**	[-2.266: -0.714]**
Adult(Ref)	- - - -	- - - -	- - - -
Colored	[-0.812: 1.332]	[-0.133: 0.221]	[-0.176: 1.339]
Indian/Asian	[-2114.690:1931.030]	[-19.123:368.214]	[-40.663: 41.971]
White	[-1170.784: 1091.783]	[-57.471: 716.553]	[-74.314: 78:948]
Black(Ref)	- - - -	- - - -	- - - -
Married	[-1.231: 0.313]	[-1.605: -0.334]	[-1.415: 0.508]
Single	[0.211: 1.106]**	[0.579: 1.540]**	[0.422: 1.250]**
Widowed	[-2005.323: 2000.254]	[-1.137: 0.679]	[-1.035: 0.755]
Divorced(Ref)	- - - -	- - - -	- - - -
Primary	[-0.796: 0.704]	[-1.040: 1.122]	[-0.801: 0.474]
Secondary	[-1.520: 0.059]	[-3.011: -1.130]**	[-1.516: 0.114]
Tertiary	[-2.745: -0.107]**	[-2.353: -0.392]**	[-2.758: -0.146]**
Otherk(Ref)	- - - -	- - - -	- - - -
Average prediction Error	0.0344	0.0345	0.0343

NB: *** Indicates the significance of the variable within [95%] H.P.D. credible interval for the coefficients of the model.

The models produced similar results regarding the association between the response variable and the explanatory variables. However, there are only few instances where the modules produced different results. The precision of credible intervals is well distinguishable for the covariates Coloured, Indian/Asian, and Single. It is evident that the modified version of Zellner's informative g-prior distribution yielded narrower credible intervals while the hyper g-prior distribution yielded wider credible intervals. The average prediction errors for the three models are similar. Even so, the modified version of Zellner's informative g-prior distribution resulted to a model with the smallest average prediction error compared to other models. For that reason, inference in this study will be base on the model developed through Zellner's g-prior distribution.

3.1.4 Results from the selected Bayesian model

Table 4: Parameter estimates for logistic regression model with an informative g-prior.

Covariates	Estimates	Std.Errors	Odds Ratio (OR)	95% HPD Credible intervals
Intercept	-6.086	2.034	0.002	[-6.778: -5.383]***
Free State	1.231	1.288	3.425	[0.179: 1.354]**
Gauteng	-3.447	1.426	0.032	[-1.637: -0.116]**
Kwa Zulu-Natal	-0.378	0.957	0.685	[-0.649: 0.534]
Limpopo	-0.720	1.649	0.478	[-1.357: 0.380]
Mpumalanga	-0.334	0.559	0.716	[-0.660: 0.732]
Nothern Cape	-0.454	1.796	0.0.635	[-1.621: 0.279]
North West	0.001	0.049	1.001	[-0.479: 1.520]
Western Cape	-1.665	3.364	0.189	[-3.418: -0.216]**
Easter Cape(Ref)	- - - -	- - - -	- - - -	- - - -
Male	1.498	1.073	4.473	[0.947: 2.197]**
Female(Ref)	- - - -	- - - -	- - - -	- - - -
Elderly	-1.1587	1.621	0.314	[-1.366: 0.481]
Youth	-1.461	1.429	0.232	[-2.266: -0.714]**
Adult(Ref)	- - - -	- - - -	- - - -	- - - -
Colored	0.033	0.166	1.034	[-0.176: 1.339]
Indian/Asian	-3.335	21.431	0.036	[-40.663: 41.971]
White	-3.205	35.018	0.041	[-74.314: 78:948]
Black(Ref)	- - - -	- - - -	- - - -	- - - -
Married	-0.392	1.820	0.676	[-1.415: 0.508]
Single	1.242	1.808	3.463	[0.422: 1.250]**
Widowed	-0.018	1.756	0.982	[-1.035: 0.755]
Divorced(Ref)	- - - -	- - - -	- - - -	- - - -
Primary	0.043	0.454	1.044	[-0.801: 0.474]
Secondary	-0.412	0.858	0.662	[-1.516: 0.114]
Tertiary	-1.264	2.429	0.283	[-2.758: -0.146]**
Otherk(Ref)	- - - -	- - - -	- - - -	- - - -

NB: *** Indicates the significance of the variable within [95%] H.P.D. credible interval for the coefficients of the model.

3.2 Discussion

When the 95 highest posterior density (HPD) credible intervals (CIs) do not include zero, the variables involved are significant. This implies a significant relationship between the response variable (alcohol and drugs misuse) and the explanatory variable involved. Among the provinces of South Africa, Free State and Western Cape are significant, as shown in Table 4. The estimate and the 95% H.P.D. credible intervals for Free State province are positive. This signifies a positive association between the province and the misuse of alcohol and drugs compared to the association between Eastern Cape (the reference category) and the misuse of alcohol and drugs. The estimate and the 95 HPD credible intervals for Western Cape Province are negative. This signifies a negative association between the province and the misuse of alcohol and drugs compared to the association between Eastern Cape (the reference category) and the misuse of alcohol and drugs.

Based on gender, the results provide evidence of a positive relationship between males and the misuse of alcohol and drugs compared to females. The age group revealed to have a negative association with drugs and alcohol misuse is youth, and there is no evidence of association between the elderly and the misuse of alcohol and drugs. This is shown in Table 4, where the reference category is adult. These results suggest that youth is less likely to misuse alcohol and drugs than adults. The results showed no evidence of an association between race and the misuse of alcohol and drugs.

Based on marital status, the results unfold a positive relationship between single people and the misuse of alcohol and drugs. In this case, the reference category is divorced people. This signifies that single people misuse alcohol and drugs more than divorced, married, and widowed. The results revealed relationship between tertiary level of education and the misuse of alcohol and drugs. This indicates that people with tertiary level of education are less likely to misuse alcohol and drugs compared to people with no educational background or primary level education.

The odds ratio (OR) denoted by e^{β} provide a basic interpretation of the magnitude of the regression model coefficient β. When the OR is greater than 1, it indicates a positive relationship between the response and explanatory variables. That is, for every one-unit increase in an explanatory variable x, the odds increase multiplicatively by e^{β}. The odds ratio of 3.425 for Free State implies that people in Free State are 3 times more likely to misuse alcohol and drugs than people in Eastern Cape. However, the odds ratio of 0.189 for the Western Cape implies that people are 18.9% less likely to misuse alcohol and drugs than people in Eastern Cape. Similarly, the odds ratio of 4.473 means that males in South Africa are 4 times more likely to misuse alcohol and drugs than females. Furthermore, the odds ratio of 3.463 implies that single people are almost 4 times more likely to misuse alcohol and drugs than divorced people. The odds ratio of 0.232 for youth reveals that there is 23.2% less chance for youth to misuse alcohol and drugs when compared to adults while he odds ratio of 0.283 for tertiary implies 28.3% less chance for people with tertiary level of education to misuse alcohol and drugs than people who did not attain a primary level of education or do not have an educational background.

This study is limited to one year survey data. As a result, the study could not explore the trends of alcohol and drugs misuse over several years. Therefore, statistical inferences made are limited to 2019 incidents. High variability in the dataset were observed on the variables Age group, Population group and Marital status. However, the high variability only affected the credible intervals precision of the variables involved. The modified version of Zellner's informative g-prior distribution used in this study is effective and can be used in other datasets where the assumptions of a binary logistic regression are not violated. Further research may explore the effectiveness of some of the recommended prior distributions for the GLMs such as Jeffreys' prior, weakly-informative Cauchy priors and other variants of Zellner's informative g-prior distributions. Further research may also include comparisons of GLMs in Bayesian framework and Machine Learning algorithms.

4 Conclusion

In this study, the use of a modified version of Zellner's informative g-prior distribution resulted to a Bayesian binary logistic regression model that provides precise parameter estimates and can predict the misuse of alcohol and drugs effectively. The model developed also accounted for separation encountered in classical binary logistic regression model. The results revealed that the population at risk of misusing alcohol and drugs include males, people with no primary level of education or no educational background and single people. Based on the findings of this study, we recommend that the South African intervention programs, such as the South African National Council on Alcohol and Drugs (SANCA), Anti-Substance Misuse Program of Action, etc., aiming to address issues of alcohol and drugs misuse develop intervention and educational programs that can cater for people with no educational background and those who did not attain the primary level of education. These programs may also target the Free State province of South Africa since it is identified as the province more cases of alcohol and drugs misuse.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgement

We appreciate Statistics South Africa for providing us with the data used in this research.

References

[1] A. T. McLellan, Substance misuse and substance use disorders: why do they matter in healthcare?, Transactions of the American Clinical and Climatological Association, WHO **128**, 112 (2017).
[2] World Health Organization, Global status report on alcohol and health 2018: executive summary (No. WHO/MSD/MSB/18.2), (2018).
[3] A. M. Laslett and R. Room and O. Waleewong, and O. Stanesby and S. Callinan and World Health Organization, Harm to others from drinking: Patterns in nine societies, WHO (2019).

[4] H. Ritchie and M. Roser, Drug use, Our World in Data, (2019).

[5] United Nations Office on Drugs and Crime, World Drug Report, United Nations Publication, Sales No. E. 18. XI. 9 (2018).

[6] B. Parker, 80 % of S.A.'s male youth deaths are alcohol-related and drug consumption is twice the world norm, Available online: https://www.parent24.com/Family/Health/80-of-sas-male-youth-deaths-are-alcohol-related-and-drug-consumption-is-twice-the-world-norm-20180626, (Accessed on 19 August 2021).

[7] S. Dada, and H.N Burnhams and J. Erasmus and L. C. W. Parry and B. S. A. Pretorius and W. H. R. Keen, Monitoring alcohol, tobacco and other drug misuse treatment admissions in South Africa, SACENDU, 1-72 (2018).

[8] World Health Organization, Global status report on alcohol and health, WHO, Thailand, (2019).

[9] Substance A. Mental H.S.A., Impact of the DSM-IV to DSM-5 Changes on the National Survey on Drug Use and Health, SAMHSA, (2016).

[10] NIDA, Brains, and Behavior: The Science of Addiction, NIDA (2014).

[11] R. Bränström and S. Andréasson, Regional differences in alcohol consumption, alcohol addiction and drug use among Swedish adults, Scandinavian Journal of Public Health, **36**(5), 493-503 (2008).

[12] P. Mafa and J.C. Makhubele and J.A. Ananias and B.N. Chilwalo and F.K. Matlakala and S.F. Rapholo and A. Svinurai and M.W. Hasheela and N.I.H. Tiberia and R.J. Freeman, Alcohol consumption patterns: A gender comparative study among high school youth in South Africa, Global Journal of Health Science, **11**(2), 92-101 (2019).

[13] Z. Mbandlwa and N. Dorasamy, The impact of substance misuse in South Africa: a case of informal settlement communities, Journal of Critical Reviews, **7**(19), (2020).

[14] D. Shmulewitz and D.S. Hasin, Risk factors for alcohol use among pregnant women, ages 15–44, in the United States, 2002 to 2017, Preventive Medicine, **124**, 75-83 (2019).

[15] B. Vythilingum and D.J. Stein and A. Roos and S.C. Faure, and L. Geerts, Risk factors for substance use in pregnant women in South Africa, South African Medical Journal, 102(11), 851-854 (2012).

[16] A. Gelman and A. Jakulin and M.G. Pittau and Y.S. Su, A weakly informative default prior distribution for logistic and other regression models, The Annals of Applied Statistics, **2**(4), 1360-1383 (2008).

[17] M. Suleiman and H. Demirhan and L. Boyd and F. Girosi and V. Aksakalli, Bayesian logistic regression approaches to predict incorrect D.R.G. assignment, Health Care Management science, **22**(2), 364-375 (2019).

[18] E. Hanson and A.J. Branscum and W.O. Johnson, Informative g-priors for logistic regression, Bayesian Analysis, **9**(3), 597-612 (2014).

[19] N.R. Lally, 2015. The Informative g-Prior vs. Common Reference Priors for Binomial Regression With an Application to Hurricane Electrical Utility Asset Damage Prediction, B.A. University of Connecticut (2015).

[20] A. Zellner, Models, prior information, and Bayesian analysis, Journal of Econometrics, **75**(1), 51-68 (1996).

[21] A. Zellner, Bayesian Methods and Entropy in Economics and Econometrics, In Fundamental Theories of Physics, W.T. Grandy and L.H. Schick, Eds., Springer: Dordrecht, Netherlands, **43**, 17-31 (1991).

[22] R.E. Kass and L. Wasserman, The selection of prior distributions by formal rules, Journal of the American Statistical Association, **91**(435), 1343-1370 (1996).

[23] D. Fouskakis and I. Ntzoufras and D. Draper, Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care, The Annals of Applied Statistics, **3**(2), 663-690 (2009).

[24] F. Liang and R. Paulo and G. Molina and M.A. Clyde and J.O.Berger, Mixtures of g priors for Bayesian variable selection, Journal of the American Statistical Association, **103**(481), 410-423 (2008).

APPENDIX

Table 5: Cross-tabulation of misuse of alcohol and drugs by Provinces.

Variable	Misuse alcohol and drugs	
	Yes	No
EC	14	6214
	0.03%	12.90%
FS	21	2857
	0.05%	5.93%
GP	13	11376
	0.03%	23.62%
KZN	21	8587
	0.05%	17.83%
LP	5	5212
	0.01%	10.82%
MP	11	3935
	0.02%	8.17%
NC	5	2141
	0.01%	4.44%
NW	6	3125
	0.01%	6.49%
WC	3	4618
	0.00%	9.59%

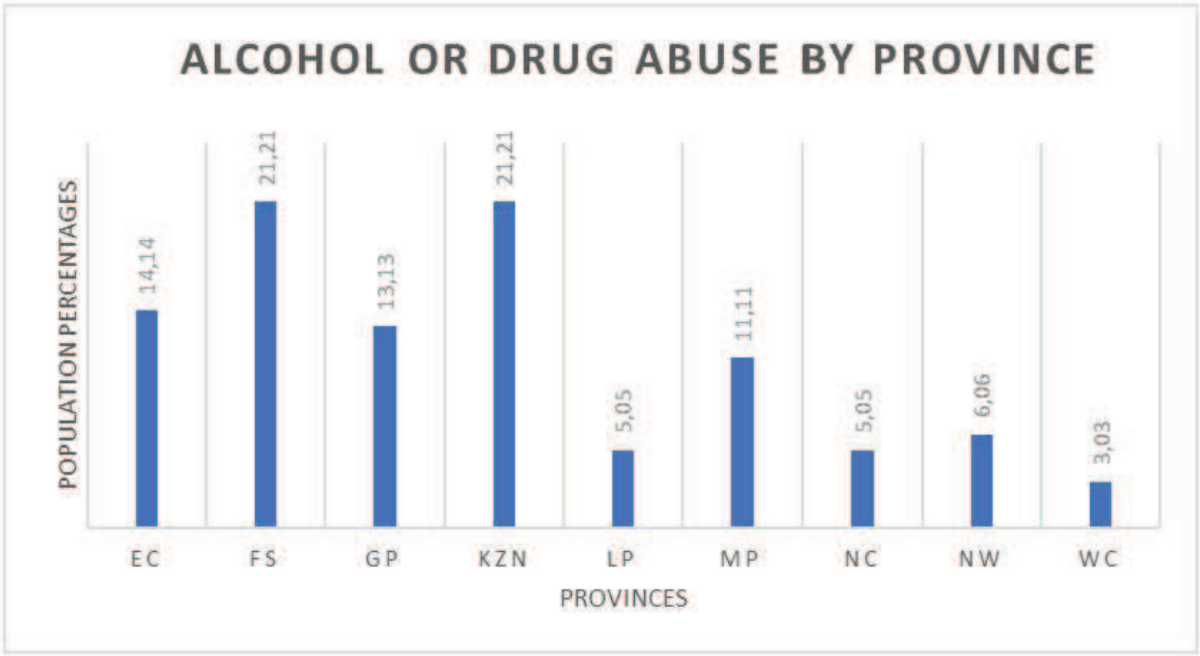

Fig. 2: The distribution of alcohol and drug misuse in South Africa by provinces.

Table 6: Cross-tabulation of misuse of alcohol and drugs by gender.

Gender	Misuse alcohol and drugs	
	Yes	No
Male	83	22446
	0.17%	46.60%
Female	16	25619
	0.03%	53.19%

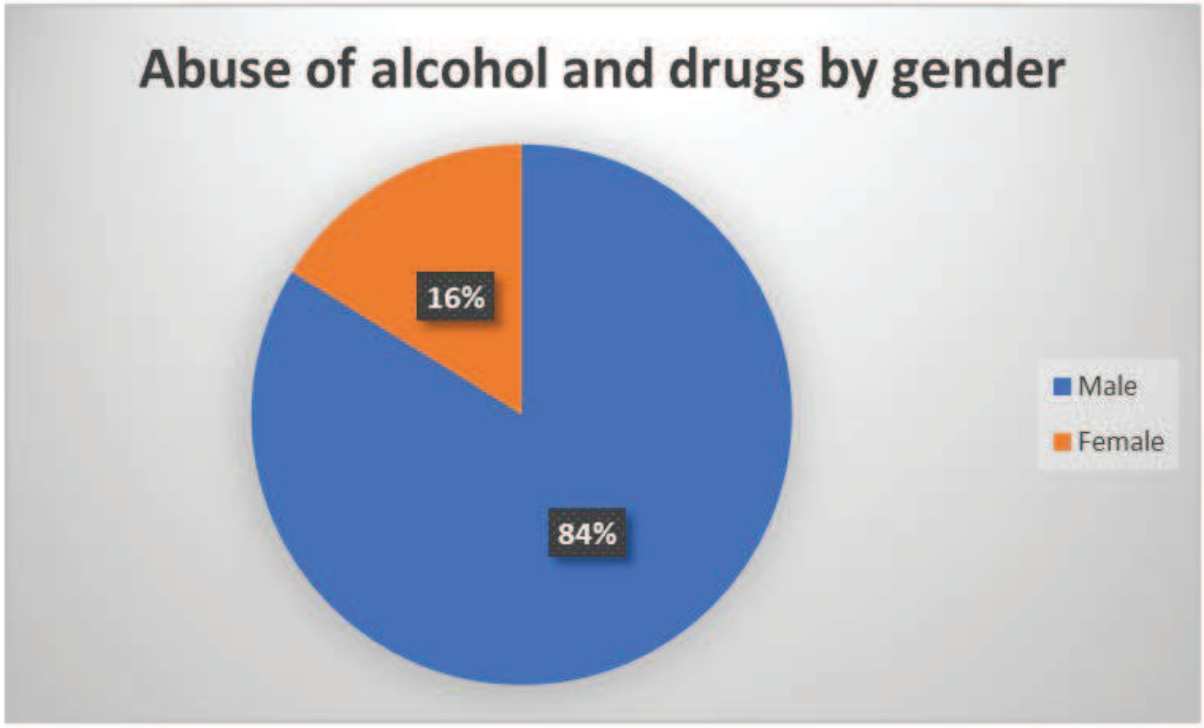

Fig. 3: The distribution of alcohol and drug misuse in South Africa by gender.

Table 7: Cross-tabulation of misuse of alcohol and drugs by Age.

Age	Misuse alcohol and drugs	
	Yes	No
Adult	85	31836
	0.18%	66.10%
Elderly	5	4642
	0.01%	9.64%
Youth	9	11587
	0.02%	24.06%

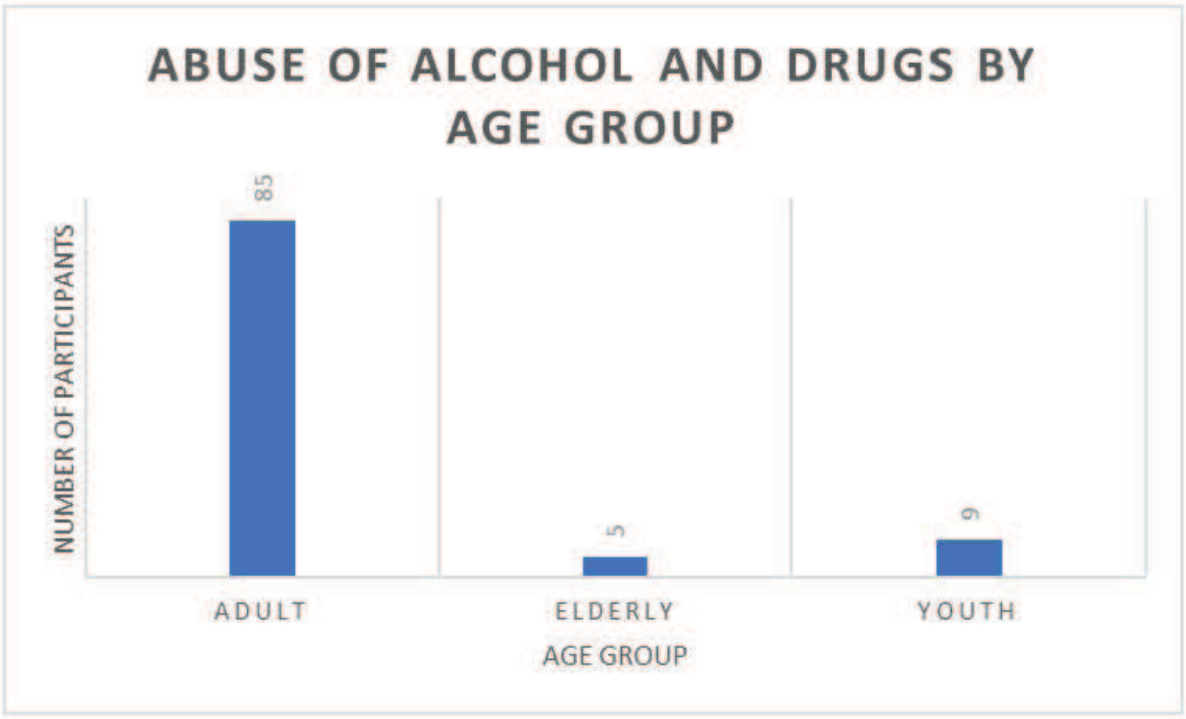

Fig. 4: The distribution of alcohol and drug misuse in South Africa by age.

Table 8: Cross-tabulation of misuse of alcohol and drugs by race.

Race	Misuse alcohol and drugs	
	Yes	**No**
African	14	6214
	0.03%	12.90%
Colored	21	2857
	0.05%	5.93%
Indian/Asian	13	11376
	0.03%	23.62%
White	21	8587
	0.05%	17.83%

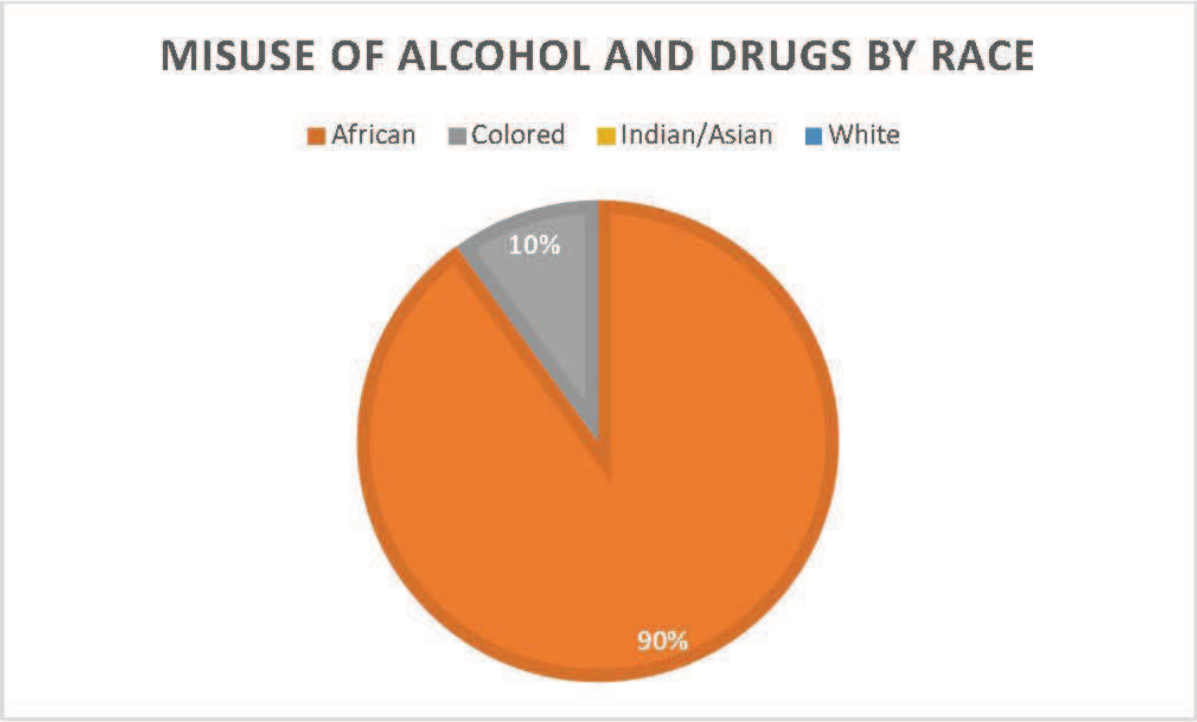

Fig. 5: The distribution of alcohol and drug misuse relative to race.

Table 9: Cross-table of alcohol and drugs misuse by Marital status.

Marital Status	**Misuse alcohol and drugs**	
	Yes	**No**
Married	18	12988
	0.04%	26.97%
Single	75	30412
	0.16%	63.14%
Divorced	0	964
	0.00%	2.00%
Widowed	6	3701
	0.01%	7.68%

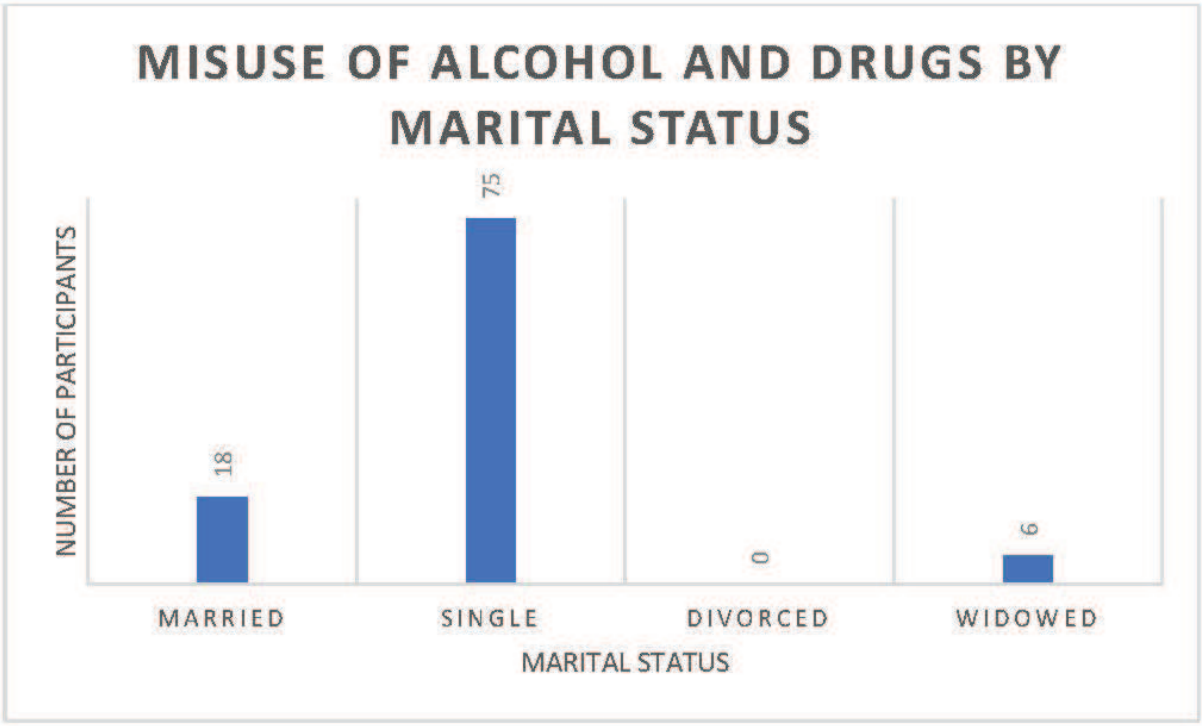

Fig. 6: The distribution of alcohol and drug misuse relative to marital status.

Table 10: Cross-table of alcohol and drugs misuse by education level.

EduLevel	**Misuse alcohol and drugs**	
	Yes	**No**
Primary	53	14729
	0.11%	30.58%
Secondary	37	24884
	0.08%	51.67%
Tertiary	3	5551
	0.00%	11.53%
Other	6	3701
	0.01%	6.02%

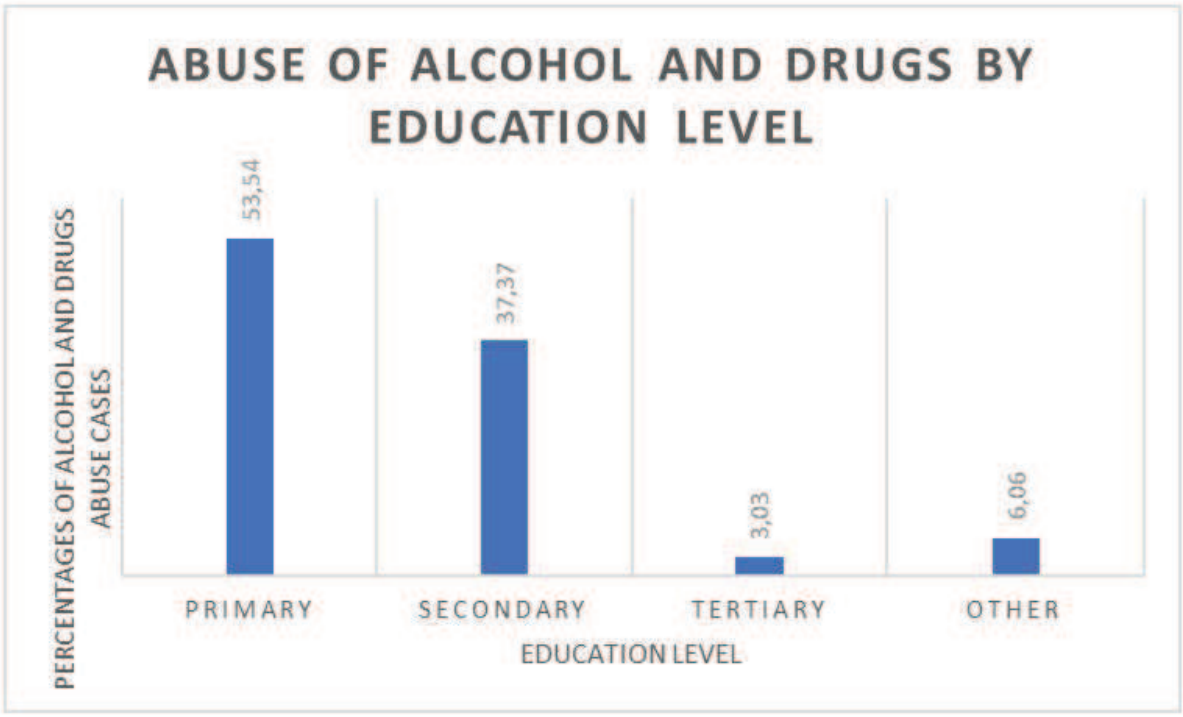

Fig. 7: The distribution of alcohol and drugs misuse relative education level.

Table 11: Detection of separation test results

Covariates	MLE value
Intercept	-inf
Free State	0
Gauteng	0
Kwa Zulu-Natal	0
Limpopo	0
Mpumalanga	0
Nothern Cape	0
North West	0
Western Cape	0
Easter Cape(Ref)	- - - -
Male	0
Female(Ref)	- - - -
Elderly	0
Youth	0
Adult(Ref)	- - - -
Colored	0
Indian/Asian	-inf
White	-inf
Black(Ref)	- - - -
Married	inf
Single	inf
Widowed	inf
Divorced(Ref)	- - - -
Primary	0
Secondary	0
Tertiary	0
Otherk(Ref)	- - - -

0:finite value, inf:infinity, and -inf:-infinity