Marijuana and Tobacco Consumption in Colombia: A Bayesian Zero-Inflated Ordered Probit Model*

Alejandro Pinilla-Barrera† Andrés Ramírez-Hassan‡

Santiago Montoya-Blandón§

Abstract

This paper develops a Bayesian implementation of the zero-inflated ordered probit model introduced in Harris and Zhao (2007), providing a broad replication of the frequentist method and extending the original paper with a Bayesian implementation. Posterior distributions naturally propagate parameter uncertainty to non-linear quantities of interest—participation, consumption intensity, and associated marginal effects—without relying on asymptotic approximations such as the delta method. Performance is assessed through Monte Carlo experiments compared against maximum likelihood estimation, with empirical applications to marijuana and tobacco consumption in Colombia. Results confirm the Bayesian approach reproduces original substantive findings while providing a coherent framework for structural inference.

Keywords: Bayesian inference; Zero-inflated ordered probit; Psychoactive substance consumption; Hamiltonian Monte Carlo.

JEL: C11; C35; C63

*We acknowledge the supercomputing resources that were made available by the Centro de Computación Científica Apolo at Universidad EAFIT (http://www.eafit.edu.co/apolo) to conduct the research reported in this paper. All remaining errors are our own. This work was supported by Universidad EAFIT.

†School of Finance, Economics and Government, Universidad EAFIT, Medellín, Colombia. E-mail: apinill2@eafit.edu.co. Carrera 49 7 Sur-50, phone: + 57 2619500.

‡School of Finance, Economics and Government, OMEGA research group, Universidad EAFIT, Medellín, Colombia. E-mail: aramir21@eafit.edu.co.

§Adam Smith Business School, University of Glasgow, Glasgow, United Kingdom. E-mail: santiago.montoya-blandon@glasgow.ac.uk.

1 Introduction

According to the 2023 United Nations Office on Drugs and Crime report (UNODC, 2023), an estimated 219 million people used marijuana in 2021, representing about 4.3% of the global population aged 15–64. Marijuana remains the most widely consumed illicit psychoactive substance worldwide, and the number of users has increased by approximately 23% over the past decade. At the same time, tobacco—a legal and widely available substance—continues to be one of the leading causes of preventable morbidity and mortality worldwide, despite substantial declines in smoking prevalence in many countries. Marijuana and tobacco consumption generate important social, health, and economic consequences, which requires understanding the determinants of participation and consumption intensity across markets.

Empirical analyses on psychoactive substance consumption frequently rely on survey data where the outcome is recorded as ordered consumption intensity categories (e.g., no use, occasional use, frequent use). A common feature of such data is a large proportion of zero observations. As emphasized by Harris and Zhao (2007), these zeros may arise from two distinct behavioral mechanisms. Some individuals are genuine non-participants in the market due to health, legal, moral, or personal reasons. Others are potential consumers who reported zero consumption but whose behavior is still governed by standard economic factors such as prices, income, and preferences. Furthermore, both behaviors can be correlated through unobserved confounders such as addiction, family background and personal history. Ignoring this distinction may lead to misleading inference and incorrect policy conclusions.

To address this issue, Harris and Zhao (2007) propose a *zero-inflated ordered probit* (ZIOP) model in which outcomes are generated by two latent processes: a binary probit for the participation decision and an ordered probit governing the intensity of consumption conditional on participation. The general framework allows correlation between the two latent processes driving zero (no consumption) outcomes.

In this paper, we develop a Bayesian inferential framework for the zero-inflated ordered probit model with correlated disturbances. Our contribution is twofold. First, we provide a broad replication of the framework introduced by Harris and Zhao (2007) by reproducing its substantive implications under an alternative inferential paradigm and benchmarking the resulting estimates against maximum likelihood estimation. We use the term broad replication to emphasize that our objective is to replicate the econometric framework and its substantive conclusions rather than to reproduce the original empirical application using the same dataset. Second, we extend the original contribution by developing a Bayesian implementation of the model.

While the original model of Harris and Zhao (2007) is estimated using maximum likelihood, the Bayesian approach offers several advantages. First, posterior simulation naturally propagates parameter uncertainty into predictions and all derived quantities of interest. Second, inference for nonlinear functions of the parameters, such as marginal effects and choice probabilities, is obtained directly from posterior draws, avoiding the need for asymptotic approximations such as the delta method or simulated asymptotic sampling. Third, the Bayesian framework delivers the full posterior distribution of structural quantities, allowing uncertainty quantification and scenario analysis to be carried out in a probabilistically coherent manner.

We illustrate the proposed approach by studying the determinants of marijuana and tobacco consumption in Colombia, a particularly relevant context as the country is among the world's major producers of marijuana (UNODC, 2021), contributing to relatively low prices and widespread availability. Moreover, as the most consumed illicit psychoactive substance, marijuana warrants particular attention, especially given that the Colombian Congress has repeatedly debated the legalization of recreational cannabis. Tobacco, a legal and heavily regulated substance, ranks second in consumption only to alcohol, with prevalence rates exceeding one third of the population, particularly among the young. This application provides a useful benchmark for comparing two markets that differ substantially in their legal status, social perceptions, and regulatory environments, while allowing a direct parallel to the original application in Harris and Zhao (2007).

The remainder of the paper is organized as follows. Section 2 outlines the econometric model and the Bayesian estimation strategy, Section 3 describes the data, Section 4 presents the empirical results and Section 5 provides conclusions alongside policy implications.

2 Model Specification

We follow the zero-inflated ordered probit (ZIOP) with correlated disturbances model proposed by Harris and Zhao (2007) (noted as ZIOPC in the original paper; see Section 2.2). The specification extends the standard ordered probit (OP) by allowing excess zeros to arise from two distinct behavioral processes: non-participation and corner solutions among potential consumers.

Let y denote an observed ordered outcome taking values among the set of J alternatives denoted as $\mathcal{J} \coloneqq \{0, 1, \ldots, J\}$. In many empirical applications, including survey data on consumption of addictive goods like marijuana or tobacco, the distribution of y displays a large mass at zero. In our case of ordered consumption frequencies, this corresponds to individuals who report no recent consumption (Harris and Zhao, 2007). Standard ordered probit models are ill-suited for such settings because they rely on a single latent process and therefore cannot distinguish between individuals who choose not to participate and those who would participate but face a corner solution resulting in zero consumption.

To address this issue, the ZIOP model combines a binary equation governing the participation decision with an ordered outcome equation governing consumption intensity. Individuals first decide whether to participate in the market given some observables, and conditional on participating, they determine their level of consumption. Zero outcomes may therefore arise either because individuals opt out of market participation altogether or because those who choose to participate maximize their latent utility by demanding zero units.

Let r denote a binary participation indicator such that $r = 1$ if an individual participates in the market and $r = 0$ otherwise. The marginal utility of joining the market is determined by a latent variable r^*, such that $r = \mathbb{I}(r^* > 0)$, where $\mathbb{I}$ is the indicator function that equals 1 when its argument is true and is zero otherwise. Conditional on participation ($r = 1$), consumption intensity is determined by an OP model with latent variable $\tilde{y}^*$. These latent processes are jointly modeled as

$$r^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon, \tag{1}$$

$$\tilde{y}^* = \boldsymbol{z}'\boldsymbol{\gamma} + u, \tag{2}$$

where $\boldsymbol{x}$ is a d_x-dimensional vector of covariates that determine participation, $\boldsymbol{z}$ is a d_z-dimensional vector of covariates that determines intensity, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are d_x- and d_z-dimensional parameter vectors, respectively, with ε containing all remaining unobservable factors influencing participation and a second set of unobservables u influencing intensity.

The observable intensity guided only be the OP process is denoted as $\tilde{y}$. Given thresholds $\boldsymbol{\mu} := (\mu_{-1}, \mu_0, \mu_1, \ldots, \mu_{J-1}, \mu_J)$, with the convention that $\mu_{-1} = -\infty$ and $\mu_J = \infty$, $\tilde{y}^*$ maps to $\tilde{y}$ according to $\tilde{y} = j \cdot \mathbb{I}(\mu_{j-1} < \tilde{y}_i^* \leq \mu_j)$ for $j = 0, \ldots, J$. Such specification also sets the first threshold to 0 ($\mu_0 = 0$) for identification purposes (equivalent to setting the ordered equation intercept $\gamma_1 = 0$). Correlation between unobservables ε and u that drive both processes is imposed via

$$\begin{bmatrix} \varepsilon \\ u \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right). \tag{3}$$

Note that having the diagonal elements of the variance-covariance matrix equal to

1 acts as an identification restriction. The case $\rho \neq 0$ allows for participation to be correlated to intensity after conditioning on observable characteristics $\boldsymbol{x}$ and $\boldsymbol{z}$, which is likely the case in many empirical applications. Additionally, as emphasized by Harris and Zhao (2007), an exclusion restriction becomes necessary to separate the effect of the two correlated latent variables r^* and $\tilde{y}^*$. In practice, this amounts to having at least one regressor in $\boldsymbol{x}$ that is not present in $\boldsymbol{z}$ (or viceversa).

Finally, observed intensity y arises from both binary participation r and ordered intensity $\tilde{y}$, such that $y = r \cdot \tilde{y}$. Thus, a zero outcome $y = 0$ may arise either because the individual does not participate ($r = 0$) or because the individual participates but reports zero consumption ($r = 1$ yet $\tilde{y} = 0$). In the context of our application, this means allowing for the decision to consume an addictive substance like marijuana or tobacco to inflate the probability of no consumption and to be correlated with frequency of use (after controlling for observables affecting each decision).

Given this setup, the implied probabilities for the observed outcome $y \in \mathcal{J}$ are

$$\Pr(y = 0 \mid \boldsymbol{z}, \boldsymbol{x}) = [1 - \Phi(\boldsymbol{x}'\boldsymbol{\beta})] + \Phi_2(\boldsymbol{x}'\boldsymbol{\beta}, -\boldsymbol{z}'\boldsymbol{\gamma}; -\rho) \tag{4}$$

$$\Pr(y = j \mid \boldsymbol{x}, \boldsymbol{z}) = \Phi_2(\boldsymbol{x}'\boldsymbol{\beta}, \mu_j - \boldsymbol{z}'\boldsymbol{\gamma}; -\rho) - \Phi_2(\boldsymbol{x}'\boldsymbol{\beta}, \mu_{j-1} - \boldsymbol{z}'\boldsymbol{\gamma}; -\rho) \tag{5}$$

$$\Pr(y = J \mid \boldsymbol{x}, \boldsymbol{z}) = \Phi_2(\boldsymbol{x}'\boldsymbol{\beta}, \boldsymbol{z}'\boldsymbol{\gamma} - \mu_{J-1}; \rho), \tag{6}$$

where $j \in \{1, \ldots, J-1\}$, $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of the standard normal distribution and $\Phi_2(\cdot, \cdot; \rho)$ denotes the CDF of the bi-variate standard normal distribution with correlation coefficient ρ. Equation (4) inflates the probability of zero implied by the standard OP model by using the zero probability of another probit model (and allowing for correlation between the two).

We assume access to an independent and identically distributed (*i.i.d.*) sample of size N for all observables, denoted as $\{(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)\}_{i=1}^N$. Collect $\boldsymbol{y} \coloneqq (y_1, \ldots, y_N)$, $\boldsymbol{X} \coloneqq (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_N')'$, and $\boldsymbol{Z} \coloneqq (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_N')'$. The likelihood function over all parameters $\boldsymbol{\theta} := (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \rho)$ is therefore

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = \prod_{i=1}^{N} \prod_{j=0}^{J} \Pr(y_i = j \mid \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta})^{\mathbb{I}(y_i=j)}. \tag{7}$$

Bayesian inference proceeds by specifying a joint prior distribution $\pi(\boldsymbol{\theta})$ over the unknown parameters and updating their posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ by an application of Bayes' rule:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) \times \pi(\boldsymbol{\theta}). \tag{8}$$

This posterior distribution contains all information about the parameters $\boldsymbol{\theta}$ that is available in the data. All that remains is the specification of prior distributions for our Bayesian analysis. For the regression coefficients, we employ $\boldsymbol{\beta} \sim \mathcal{N}_{d_x}(\mathbf{0}_{d_x}, \sigma^2_\beta \cdot \boldsymbol{I}_{d_x})$ and $\boldsymbol{\gamma} \sim \mathcal{N}_{d_z}(\mathbf{0}_{d_z}, \sigma^2_\gamma \cdot \boldsymbol{I}_{d_z})$, where $\mathbf{0}_p$ and $\boldsymbol{I}_p$ correspond to the zero vector and identity matrix of order p, respectively. These normal priors become diffuse when large prior variances are used (in our implementation, we set $\sigma^2_\beta = \sigma^2_\gamma = 1000$). To impose the ordering of the thresholds $\boldsymbol{\mu}$, we adopt the reparameterization of Albert and Chib (2001), namely $\delta_j = \log(\mu_j - \mu_{j-1})$ for $j = 1, \ldots, J-1$, and assign a non-informative normal prior to $\boldsymbol{\delta} := (\delta_1, \ldots, \delta_{J-1})$. In our application, we set $\boldsymbol{\delta} \sim \mathcal{N}_{J-1}(\mathbf{0}_{J-1}, 5 \cdot \boldsymbol{I}_{J-1})$. For the correlation parameter ρ, we set a fully non-informative uniform prior on its support $[-1, 1]$. For additional details of our Bayesian implementation, see Appendix A. For Monte Carlo simulations supporting the introduced Bayesian implementation, see Appendix B

3 Data

The data comes from the National Survey on the Consumption of Psychoactive Substances (ENCSPA, for its acronym in Spanish), a nationally representative survey conducted in 2019 by the Colombian national statistics office (DANE). The survey's final sample size is 45,516 respondents, representative of 23.6 million individuals or approximately 75% of the total Colombian population between 18 and 65 years old in 2019. The main objective of the survey is to measure both legal and illegal psychoactive substance abuse within the population. Households were sampled from Colombian municipalities, wherein individuals between 12 and 65 years old were randomly selected. The enumerators perform the survey privately, following up on those absent during the visit until obtaining a response; changing the chosen individual was not allowed.

Definitions of all variables used in this study are based on Harris and Zhao (2007) and presented in Table C.1 of the Appendix. In particular, an individual i's marijuana or tobacco consumption intensity levels are recorded as: zero consumption if never used ($y_i = 0$); low consumption if used previously but not in the last 12 months ($y_i = 1$); moderate if used in the last 12 months, but not in the last 30 days ($y_i = 2$); and high if used in the last 30 days ($y_i = 3$). Summary statistics for observed intensity levels of marijuana and tobacco consumption are presented in Table 1. Additional summary statistics are found in Table C.2.

Table 1: Summary of marijuana and tobacco consumption intensities

Intensities	**Marijuana**		**Tobacco**	
	Sample size	**Perc. (%)**	**Sample size**	**Perc. (%)**
No consumption	42,030	92.34%	29,619	65.07%
Low frequency	2,551	5.60%	10,286	22.60%
Moderate frequency	321	0.71%	1,052	2.31%
High frequency	614	1.35%	4,559	10.02%
Total	45,516	100%	45,516	100%

Harris and Zhao (2007) argue the participation decision is driven by individuals' attitudes toward consumption and health concerns. Variables such as curiosity about use, access to the substance, education, and socio-demographic characteristics (socio-economic strata, marital status, age, and gender) are relevant determinants of market participation. Additionally, the literature on drug consumption highlights two recurring findings: rising consumption among the young, especially young men (Pudney, 2004; Järvinen and Demant, 2011; Rotermann, 2020; Kilmer et al., 2022; Mennis et al., 2023), and the "gateway hypothesis", which posits spillover effects from soft to hard drugs, such as alcohol and tobacco use predicting marijuana consumption, or marijuana predicting cocaine and ecstasy use (Kandel, 1975; Kandel et al., 1992; DeSimone, 1998; Guttmannova et al., 2021; Kim et al., 2021; Weinberger et al., 2022). To account for these effects, we include dummy variables for alcohol/tobacco use and for young men (aged under 25) in the marijuana equations.

Conditional on participation, consumption intensity is typically modeled following a standard consumer demand framework, incorporating variables such as income and prices (Harris and Zhao, 2007). Since the ENCSPA survey does not include income information, we use socio-economic strata as a proxy for the income effect. For the marijuana application, we also include a unit price measure, computed as the ratio of monthly expenditure to average units consumed. For tobacco, as the ENCSPA does not provide price information, we incorporate province-level fixed effects to capture price heterogeneity across regions.[1]

4 Results

We perform inference of the ZIOP model with correlated disturbances using our Bayesian implementation. We run 15,000 iterations each with a burn-in of 5,000, obtaining 10,000 effective posterior draws.[2] The results are presented as marginal effects on the choice probabilities in Tables 2 and 3 for the marijuana and tobacco applications, respectively. As benchmark, we include the original Harris and Zhao (2007) frequentist maximum likelihood estimates. The model allows us to decompose the overall marginal effect on $P(y = 0)$ into two parts: the effect of non-participation $P(r = 0)$ and the effect of participation with zero consumption $P(r = 1, \tilde{y} = 0)$.

[1] Colombian tobacco prices are set by each of the 32 provinces and the capital district of Bogotá.
[2] Convergence diagnostics for all chains provided in Appendix E.1.

In general, the Bayesian and frequentist approaches produce highly comparable marginal effect estimates. The signs, magnitudes, and inferential conclusions are broadly consistent across methods, indicating that the Bayesian implementation successfully replicates the main findings obtained from maximum likelihood estimation. Although the estimated measures of uncertainty differ slightly across approaches, these differences do not materially affect the empirical conclusions.

Finally, we find that our results are broadly consistent with those of Harris and Zhao (2007), despite the fact they focus on the Australian tobacco market, whereas we consider the Colombian marijuana and tobacco markets. Nevertheless, two notable differences emerge. First, Harris and Zhao (2007) focus on young women, whereas we focus on young men, as this group is particularly relevant in the consumption of psychoactive substances in Colombia (Camacho et al., 2011). Second, while Harris and Zhao (2007) report a negative correlation parameter with considerable uncertainty around zero, our estimates provide evidence of statistically significant negative correlation for marijuana and positive for tobacco. These findings suggest that the relationship between latent participation and consumption-intensity decisions may differ both across countries and substance markets. A detailed exploration of all marginal effects by substance (marijuana and tobacco) and by equation (participation and intensity), as well as the underlying ZIOP coefficients and resulting model fit is found in Appendix D.

5 Conclusions and Policy Implications

In this paper, we develop a Bayesian implementation of the Zero-Inflated Ordered Probit model with correlated disturbances proposed by Harris and Zhao (2007). Using Hamiltonian Monte Carlo, we show that Bayesian estimation provides a practical and flexible framework for inference in the ZIOP model, allowing uncertainty to be propagated naturally to nonlinear quantities of interest such as choice proba-

Table 2: Marginal effects for marijuana consumption

Variable	ZIOP (ML)						ZIOP (Bayesian)					
	$\Pr(r=0)$	$\Pr(r=1,\tilde{y}=0)$	$\Pr(y=0)$	$\Pr(y=1)$	$\Pr(y=2)$	$\Pr(y=3)$	$\Pr(r=0)$	$\Pr(r=1,\tilde{y}=0)$	$\Pr(y=0)$	$\Pr(y=1)$	$\Pr(y=2)$	$\Pr(y=3)$
Age	3.974%	-3.207%	0.767%	-0.740%	-0.016%	-0.010%	3.431%	-2.584%	0.847%	-0.784%	-0.040%	-0.023%
	(0.923%)*	(0.907%)*	(0.134%)*	(0.126%)*	(0.006%)*	(0.005%)*	(1.440%)*	(1.205%)*	(0.270%)*	(0.310%)*	(0.038%)	(0.012%)
Strata												
High	-4.901%	3.872%	-1.029%	0.989%	0.024%	0.016%	-4.374%	3.280%	-1.094%	1.026%	0.042%	0.026%
	(1.847%)*	(1.691%)*	(0.310%)*	(0.295%)*	(0.011%)*	(0.009%)	(2.349%)	(1.961%)	(0.475%)*	(0.473%)*	(0.021%)*	(0.012%)*
Medium	-1.138%	0.859%	-0.279%	0.268%	0.007%	0.005%	-1.005%	0.701%	-0.304%	0.277%	0.016%	0.011%
	(0.457%)*	(0.424%)*	(0.074%)*	(0.070%)*	(0.003%)*	(0.003%)	(0.555%)	(0.467%)	(0.113%)*	(0.125%)*	(0.014%)	(0.007%)
Years of education	0.162%	-0.134%	0.028%	-0.027%	-0.001%	0.000%	0.141%	-0.110%	0.031%	-0.029%	-0.001%	-0.001%
	(0.053%)*	(0.050%)*	(0.007%)*	(0.007%)*	(0.000%)*	(0.000%)	(0.070%)*	(0.059%)	(0.012%)*	(0.013%)*	(0.001%)*	(0.000%)
Male	-1.573%	0.600%	-0.973%	0.900%	0.040%	0.033%	-1.398%	0.427%	-0.971%	0.852%	0.060%	0.059%
	(0.438%)*	(0.382%)*	(0.123%)*	(0.112%)*	(0.012%)*	(0.014%)*	(0.600%)*	(0.362%)	(0.313%)*	(0.327%)*	(0.012%)*	(0.017%)*
Children	-0.810%	0.875%	0.065%	-0.054%	-0.006%	-0.005%	-0.689%	0.736%	0.046%	-0.029%	-0.009%	-0.009%
	(0.432%)	(0.398%)*	(0.068%)	(0.066%)	(0.003%)*	(0.003%)	(0.478%)	(0.448%)	(0.074%)	(0.070%)	(0.003%)*	(0.004%)*
Married	-0.249%	0.458%	0.209%	-0.199%	-0.006%	-0.004%	-0.179%	0.394%	0.215%	-0.195%	-0.011%	-0.009%
	(0.480%)	(0.452%)	(0.051%)*	(0.048%)*	(0.003%)*	(0.002%)	(0.438%)	(0.423%)	(0.085%)*	(0.089%)*	(0.005%)*	(0.003%)*
Capital	-0.989%	0.604%	-0.385%	0.364%	0.012%	0.009%	-0.857%	0.460%	-0.397%	0.363%	0.018%	0.016%
	(0.433%)*	(0.394%)	(0.083%)*	(0.078%)*	(0.005%)*	(0.005%)	(0.500%)	(0.379%)	(0.157%)*	(0.154%)*	(0.004%)*	(0.004%)*
Young man	2.080%	-1.882%	0.198%	-0.200%	0.000%	0.002%	1.769%	-1.527%	0.242%	-0.241%	-0.003%	0.002%
	(0.574%)*	(0.576%)*	(0.067%)*	(0.065%)*	(0.002%)	(0.002%)	(0.818%)*	(0.742%)*	(0.104%)*	(0.106%)*	(0.006%)	(0.004%)
Alcohol/Tobacco use	-3.113%	2.588%	-0.525%	0.508%	0.010%	0.006%	-2.682%	2.082%	-0.600%	0.567%	0.020%	0.013%
	(0.741%)*	(0.770%)*	(0.095%)*	(0.089%)*	(0.004%)*	(0.004%)	(1.164%)*	(0.996%)*	(0.208%)*	(0.214%)*	(0.007%)*	(0.004%)*
Access marijuana	-2.236%	1.808%	-0.428%	0.414%	0.009%	0.006%	-1.937%	1.456%	-0.481%	0.452%	0.017%	0.012%
	(0.511%)*	(0.524%)*	(0.075%)*	(0.069%)*	(0.004%)*	(0.003%)	(0.823%)*	(0.686%)*	(0.164%)*	(0.171%)*	(0.007%)*	(0.004%)*
Curiosity	-55.585%	24.920%	-30.665%	26.504%	1.968%	2.193%	-50.284%	21.622%	-28.662%	23.601%	2.286%	2.774%
	(3.466%)*	(3.714%)*	(1.206%)*	(0.945%)*	(0.204%)*	(0.339%)*	(15.737%)*	(8.481%)*	(7.720%)*	(8.499%)*	(0.422%)*	(0.558%)*
Price		0.169%	0.169%	-0.159%	-0.006%	-0.004%		0.166%	0.166%	-0.151%	-0.007%	-0.008%
		(0.036%)*	(0.036%)*	(0.034%)*	(0.002%)*	(0.002%)		(0.071%)*	(0.071%)*	(0.065%)*	(0.006%)	(0.002%)*
Threshold 1	1.384						1.223					
	(0.071)*						(0.650)*					
Threshold 2	1.661						1.809					
	(0.078)*						(0.390)*					
Correlation	-0.551						-0.463					
	(0.072)*						(0.079)*					

Notes: Marginal effects for non-participation and zero consumption for the marijuana application of the ZIOP model with correlated disturbances by the frequentist and Bayesian approaches. Statistically significant effects marked with $*$ based on 95% credible intervals for Bayesian methods and the significant test for the frequentist.

Table 3: Marginal effects for tobacco consumption

Variable	ZIOP (ML)						ZIOP (Bayesian)					
	$\Pr(r=0)$	$\Pr(r=1,\tilde{y}=0)$	$\Pr(y=0)$	$\Pr(y=1)$	$\Pr(y=2)$	$\Pr(y=3)$	$\Pr(r=0)$	$\Pr(r=1,\tilde{y}=0)$	$\Pr(y=0)$	$\Pr(y=1)$	$\Pr(y=2)$	$\Pr(y=3)$
Age	-25.594%	17.454%	-8.140%	6.586%	0.412%	1.141%	-26.573%	17.940%	-8.633%	6.956%	0.448%	1.229%
	(2.721%)*	(2.456%)*	(1.277%)*	(0.884%)*	(0.130%)*	(0.403%)*	(2.778%)*	(2.365%)*	(1.097%)*	(0.816%)*	(0.075%)*	(0.291%)*
Strata												
High	-11.109%	3.126%	-7.983%	4.950%	0.570%	2.463%	-11.430%	3.386%	-8.045%	5.017%	0.571%	2.456%
	(3.594%)*	(3.340%)	(1.252%)*	(0.825%)*	(0.166%)*	(0.580%)*	(3.605%)*	(3.345%)	(1.284%)*	(0.833%)*	(0.107%)*	(0.587%)*
Medium	-5.525%	2.409%	-3.116%	2.092%	0.208%	0.817%	-5.516%	2.421%	-3.095%	2.083%	0.206%	0.807%
	(1.671%)*	(1.569%)	(0.535%)*	(0.361%)*	(0.065%)*	(0.226%)*	(1.698%)*	(1.592%)	(0.519%)*	(0.356%)*	(0.043%)*	(0.224%)*
Years of education	0.082%	0.757%	0.840%	-0.429%	-0.072%	-0.339%	0.052%	0.777%	0.829%	-0.417%	-0.071%	-0.341%
	(0.198%)	(0.179%)*	(0.063%)*	(0.049%)*	(0.016%)*	(0.032%)*	(0.194%)	(0.178%)*	(0.058%)*	(0.047%)*	(0.005%)*	(0.024%)*
Male	-16.795%	-4.995%	-21.790%	10.479%	1.746%	9.566%	-16.940%	-4.870%	-21.810%	10.489%	1.745%	9.576%
	(1.598%)*	(1.583%)*	(0.560%)*	(0.476%)*	(0.389%)*	(0.617%)*	(1.631%)*	(1.541%)*	(0.519%)*	(0.405%)*	(0.069%)*	(0.498%)*
Children	-3.314%	1.607%	-1.707%	1.183%	0.110%	0.413%	-3.272%	1.561%	-1.711%	1.183%	0.111%	0.417%
	(1.884%)	(1.743%)	(0.585%)*	(0.408%)*	(0.054%)*	(0.230%)	(1.907%)	(1.741%)	(0.574%)*	(0.407%)*	(0.045%)*	(0.224%)
Married	-4.903%	9.271%	4.368%	-2.083%	-0.405%	-1.880%	-4.922%	9.228%	4.306%	-2.016%	-0.402%	-1.888%
	(1.732%)*	(1.636%)*	(0.529%)*	(0.391%)*	(0.095%)*	(0.242%)*	(1.719%)*	(1.636%)*	(0.462%)*	(0.368%)*	(0.038%)*	(0.187%)*
Capital	-0.900%	-5.084%	-5.985%	2.751%	0.522%	2.712%	-1.092%	-4.902%	-5.994%	2.760%	0.520%	2.714%
	(1.586%)	(1.580%)*	(0.951%)*	(0.582%)*	(0.138%)*	(0.439%)*	(1.601%)	(1.612%)*	(0.919%)*	(0.547%)*	(0.079%)*	(0.411%)*
Young man	3.547%	-4.721%	-1.174%	0.151%	0.146%	0.876%	2.818%	-4.218%	-1.401%	0.326%	0.157%	0.918%
	(2.305%)	(2.003%)*	(0.852%)	(0.580%)	(0.074%)*	(0.346%)*	(2.327%)	(2.056%)*	(0.829%)	(0.574%)	(0.065%)*	(0.347%)*
Alcohol	-57.114%	32.908%	-24.206%	18.415%	1.411%	4.379%	-57.024%	32.769%	-24.254%	18.320%	1.430%	4.505%
	(2.492%)*	(2.911%)*	(1.076%)*	(0.888%)*	(0.302%)*	(0.464%)*	(2.553%)*	(2.717%)*	(0.873%)*	(0.523%)*	(0.086%)*	(0.353%)*
Threshold 1	1.019						1.020					
	(0.017)*						(0.018)*					
Threshold 2	1.164						1.166					
	(0.031)*						(0.019)*					
Correlation	0.365						0.348					
	(0.130)*						(0.055)*					

Notes: Marginal effects for non-participation and zero consumption for the tobacco application of the ZIOP model with correlated disturbances by the frequentist and Bayesian approaches. Statistically significant effects marked with $*$ based on 95% credible intervals for Bayesian methods and the significant test for the frequentist.

bilities and marginal effects. Monte Carlo experiments confirm that the estimator recovers population parameters well in finite samples, particularly when intercept specification across equations is carefully considered to avoid weak identification.

Applying the model to marijuana and tobacco consumption in Colombia, the Bayesian and maximum-likelihood estimates lead to similar substantive conclusions, providing additional evidence of the robustness of the original framework. Notably, the correlation structure differs markedly between substances: it is positive for tobacco, suggesting unobserved characteristics associated with participation also drive intensive consumption, while it is negative for marijuana, suggesting that participation may be more closely linked to experimentation or curiosity.

The empirical results carry several policy-relevant insights, particularly for marijuana, which is currently decriminalized but not legally commercialized in Colombia. While our analysis is not designed to identify causal policy effects, it provides evidence on factors associated with participation and consumption intensity that may inform ongoing legalization debates, which typically encompass crime, consumption, and public health concerns. Tobacco's declining prevalence under legal regulation and heavy taxation offers a useful benchmark for anticipating the longer-run effects of marijuana legalization.

These insights organize naturally around four dimensions. First, and most strikingly, curiosity about marijuana use is associated with an increase of approximately 50 percentage points in participation probability, suggesting that attitudes, perceptions, and social norms are the primary drivers of participation. Public awareness and education campaigns may deserve particular attention in policy discussions. Second, access to marijuana is associated with an increase of approximately 2 percentage points in the probability of participation, a modest effect possibly reflecting Colombia's long-standing decriminalization. Third, prior alcohol or tobacco use is associated with an increase of roughly 3 percentage points in participation probability, consistent with the gateway hypothesis, though the magnitude remains small.

Fourth, although higher prices are associated with lower consumption probabilities across all frequency categories, the effects are economically small, suggesting price-based interventions alone may have limited influence, though taxation may still play an important role in financing prevention and public health policies.

Beyond the present applications, several methodological avenues for future research remain: these include extending the framework to incorporate additional decision stages, richer dynamic specifications to study transitions between participation and intensive consumption, or more flexible ZIOP specifications involving hierarchical structures, latent heterogeneity, or alternative error distributions.

References

Albert, J. H. and Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics*, 57(3):829–836.

Camacho, A., Gaviria, A., and Rodríguez, C. (2011). *Políticas antidroga en Colombia: éxitos, fracasos y extravíos*. Universidad de los Andes.

DeSimone, J. (1998). Is marijuana a gateway drug? *Eastern Economic Journal*, 24(2):149–164.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, 195(2):216–222.

Geweke, J. (1992). [statistics, probability and chaos]: Comment: Inference and prediction in the presence of uncertainty and determinism. *Statistical Science*, 7(1):94–101.

Guttmannova, K., Fleming, C. B., Rhew, I. C., Alisa Abdallah, D., Patrick, M. E., Duckworth, J. C., and Lee, C. M. (2021). Dual trajectories of cannabis and alcohol use among young adults in a state with legal nonmedical cannabis. *Alcoholism: clinical and experimental research*, 45(7):1458–1467.

Harris, M. N. and Zhao, X. (2007). A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *Journal of Econometrics*, 141(2):1073–1099.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Järvinen, M. and Demant, J. (2011). The normalisation of cannabis use among young people: Symbolic boundary work in focus groups. *Health, risk & society*, 13(2):165–182.

Kandel, D. (1975). Stages in adolescent involvement in drug use. *Science*, 190(4217):912–914.

Kandel, D. B., Yamaguchi, K., and Chen, K. (1992). Stages of progression in drug involvement from adolescence to adulthood: further evidence for the gateway theory. *Journal of studies on alcohol*, 53(5):447–457.

Kilmer, J. R., Rhew, I. C., Guttmannova, K., Fleming, C. B., Hultgren, B. A., Gilson, M. S., Cooper, R. L., Dilley, J., and Larimer, M. E. (2022). Cannabis use among young adults in washington state after legalization of nonmedical cannabis. *American journal of public health*, 112(4):638–645.

Kim, J. H., Weinberger, A. H., Zhu, J., Barrington-Trimis, J., Wyka, K., and Goodwin, R. D. (2021). Impact of state-level cannabis legalization on poly use of alcohol and cannabis in the united states, 2004–2017. *Drug and alcohol dependence*, 218:108364.

Mennis, J., McKeon, T. P., and Stahler, G. J. (2023). Recreational cannabis legalization alters associations among cannabis use, perception of risk, and cannabis use disorder treatment for adolescents and young adults. *Addictive behaviors*, 138:107552.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. Chapman and Hall/CRC.

Pudney, S. (2004). Keeping off the grass? an econometric model of cannabis consumption in britain. *Journal of Applied Econometrics*, 19(4):435–453.

Rotermann, M. (2020). What has changed since cannabis was legalized? *Health reports*, 31(2):11–20.

Stan Development Team (2025). RStan: the R interface to Stan. R package version 2.32.7.

Stan Development Team (2026). Stan Reference Manual. Version 2.38.

UNODC (2021). World Drug Report. Technical report, United Nations publication.

UNODC (2023). World Drug Report 2023. Technical report, United Nations publication.

Weinberger, A. H., Wyka, K., and Goodwin, R. D. (2022). Impact of cannabis legalization in the united states on trends in cannabis use and daily cannabis use among individuals who smoke cigarettes. *Drug and alcohol dependence*, 238:109563.

Appendix

A Bayesian Implementation

We implement Bayesian inference using Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 1996). HMC leverages concepts from physics, specifically Hamiltonian mechanics, to construct efficient proposals for a Markov chain Monte Carlo (MCMC) sampling scheme. The system is described by the *position* variable denoting the object of interest, $\boldsymbol{\theta}$, and an auxiliary *momentum* variable, $\boldsymbol{\eta}$ (with same dimension as $\boldsymbol{\theta}$, given by $d \coloneqq d_x + d_z + J$). As is standard, we assume momentum follows a zero-mean normal distribution with mass matrix $\boldsymbol{M}$; i.e., $\boldsymbol{\eta} \sim \mathcal{N}_d(\mathbf{0}_d, \boldsymbol{M})$.

Implementation of HMC requires solving Hamilton's differential equations describing a dynamic system with potential energy $U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ and kinetic energy $K(\boldsymbol{\eta}) = \boldsymbol{\eta}'\boldsymbol{M}^{-1}\boldsymbol{\eta}/2$, which typically lack closed-form solutions. Numerical integration techniques are used to solve the discretized problem, with standard HMC employing the *leapfrog integrator*, which requires tuning parameters L (number of leapfrog steps) and ϵ (step size). The pair $(\boldsymbol{\theta}, \boldsymbol{\eta})$ is updated via a Metropolis step that preserves the target distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) \times \mathcal{N}_d(\boldsymbol{\eta} \mid \mathbf{0}_d, \boldsymbol{M})$ (Neal, 2011). It is common to assume a diagonal mass matrix $\boldsymbol{M}$, leading to an additional d tuning parameters for the algorithm.

For computational implementation, we use `Stan` (Stan Development Team, 2026) through the `R` software via the `rstan` package (Stan Development Team, 2025). `Stan` implements HMC through the No-U-Turn Sampler (NUTS), which adaptively determines the trajectory length for the leapfrog integrator —avoiding the need to pre-specify L as well as providing an estimate for ϵ— and stops each step automatically when the samples imply a reversal in position space (for further details, see Hoffman et al., 2014). During the warm-up or burn-in phase (where samples are discarded), `Stan` also tunes the step size ϵ and the diagonal entries in mass matrix $\boldsymbol{M}$

to improve sampling efficiency. Algorithm A.1 presents the basic HMC mechanism underlying this class of samplers.

Algorithm A.1 Hamiltonian Monte Carlo

Initialize $\boldsymbol{\theta}^{(0)}$ in the support of $\pi(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$
Set step size ϵ, number of leapfrog steps L, mass matrix $\boldsymbol{M}$, and total iterations S
for $s = 1, \dots, S$ **do**
    Draw $\boldsymbol{\eta}^{(s-1)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{M})$
    Set $(\boldsymbol{\theta}^c, \boldsymbol{\eta}^c) \leftarrow (\boldsymbol{\theta}^{(s-1)}, \boldsymbol{\eta}^{(s-1)})$
    $\boldsymbol{\eta}^c \leftarrow \boldsymbol{\eta}^c + (\epsilon/2) \cdot \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}^c \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$
    for $l = 1, \dots, L$ **do**
        $\boldsymbol{\theta}^c \leftarrow \boldsymbol{\theta}^c + \epsilon \cdot \boldsymbol{M}^{-1} \boldsymbol{\eta}^c$
        if $l < L$ **then**
            $\boldsymbol{\eta}^c \leftarrow \boldsymbol{\eta}^c + \epsilon \cdot \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}^c \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$
        end if
    end for
    $\boldsymbol{\eta}^c \leftarrow \boldsymbol{\eta}^c + (\epsilon/2) \cdot \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}^c \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$
    Compute

$$\alpha = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}^c \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) \mathcal{N}_d(\boldsymbol{\eta}^c \mid \mathbf{0}_d, \boldsymbol{M})}{\pi(\boldsymbol{\theta}^{(s-1)} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) \mathcal{N}_d(\boldsymbol{\eta}^{(s-1)} \mid \mathbf{0}_d, \boldsymbol{M})} \right\}$$

    Draw $U \sim \text{Uniform}(0, 1)$
    if $U \leq \alpha$ **then**
        $\boldsymbol{\theta}^{(s)} \leftarrow \boldsymbol{\theta}^c$
    else
        $\boldsymbol{\theta}^{(s)} \leftarrow \boldsymbol{\theta}^{(s-1)}$
    end if
end for

B Monte Carlo Simulations

We conduct a Monte Carlo experiment to evaluate the sampling properties of the proposed Bayesian ZIOP estimator, following the simulation scheme in Harris and Zhao (2007). We generate $N = 1{,}000$ observations from a ZIOP data-generating process in which the disturbances of the participation and intensity equations are positively correlated with $\rho = 0.5$. The remaining parameter values are set to $\boldsymbol{\gamma} = (1, -0.25, -1)$, $\boldsymbol{\beta} = (1, -0.5)$, $\mu_1 = 4.5$ and $\mu_2 = 5.5$.

For the participation equation we generate $\boldsymbol{X} = (x_1, x_2, x_3)$, where x_1 is a vector of ones (intercept), $x_2 = \ln(u)$ for $u \sim$ Uniform$(0, 100)$, and $x_3 \sim$ Bernoulli(0.4). For the intensity equation we generate $\boldsymbol{Z} = (x_2, z_2)$, where $z_2 \sim N(0, 1)$. The covariates are chosen to mimic features of the empirical application, with some regressors appearing in both the participation and intensity equations, and other regressors satisfying the exclusion restriction. These values generate an inflated zero probability of about 75% on average.

The data-generating process is replicated $R = 100$ times. For each replication we compute the posterior median as the point estimator. Table B.1 reports bias, mean absolute error (MAE), root mean squared error (RMSE), and the coverage probability of the 95% highest posterior density (HPD) intervals associated with the posterior median. Overall, the results indicate that the proposed Bayesian estimator performs well in finite samples and accurately captures the population values. Bias is small across all parameters, and the coverage probabilities of the HPD intervals are close to the nominal 95% level.

Table B.1: Monte Carlo Results: Bayesian Estimation of the Zero-Inflated Ordered Probit Model

Parameter	**True value**	**Bias**	**MAE**	**RMSE**	**Coverage**
γ_1	1.00	0.034	0.154	0.154	97%
γ_2	-0.25	-0.006	0.040	0.040	98%
γ_3	-1.00	-0.013	0.073	0.073	95%
β_2	1.00	0.049	0.109	0.109	92%
β_3	-0.50	-0.014	0.059	0.059	95%
α_1	4.50	0.083	0.378	0.378	95%
α_2	5.50	0.119	0.398	0.398	95%
ρ	0.50	-0.094	0.137	0.137	98%

Notes: Based on $R = 100$ Monte Carlo replications with sample size $N = 1{,}000$. Point estimates are posterior medians. Coverage refers to the empirical coverage of 95% HPD intervals. The model specification includes an intercept in the participation equation but not in the intensity equation.

The regression coefficients in both the participation and intensity equations are estimated with relatively small MAE and RMSE. As expected in ordered response

Table B.2: Monte Carlo Results: Bayesian Zero-Inflated Ordered Probit Model (Both Intercepts)

Parameter	**True value**	**Bias**	**MAE**	**RMSE**	**Coverage**
γ_1	1.00	508.098	508.142	508.142	43%
γ_2	-0.25	432.621	432.684	432.684	81%
γ_3	-1.00	89.206	243.030	243.030	96%
β_1	0.50	-0.486	0.658	0.658	32%
β_2	1.00	-0.654	0.690	0.690	31%
β_3	-0.50	0.264	0.286	0.286	32%
α_1	4.50	-2.583	2.803	2.803	32%
α_2	5.50	-2.910	3.135	3.135	32%
ρ	0.50	-0.309	0.353	0.353	96%

Notes: Based on $R = 100$ Monte Carlo replications with sample size $N = 1{,}000$. Point estimates are posterior medians. Coverage refers to the empirical coverage of 95% HPD intervals. The model specification includes intercepts in both the participation and ordered equations.

Table B.3: Monte Carlo Results: Bayesian Zero-Inflated Ordered Probit Model (No Participation Intercept)

Parameter	**True value**	**Bias**	**MAE**	**RMSE**	**Coverage**
γ_2	-0.25	0.134	0.145	0.145	94%
γ_3	-1.00	6.610	6.725	6.725	97%
β_1	0.50	0.153	0.600	0.600	98%
β_2	1.00	0.054	0.154	0.154	97%
β_3	-0.50	-0.024	0.116	0.116	93%
α_1	4.50	0.025	0.731	0.731	98%
α_2	5.50	0.092	0.792	0.792	98%
ρ	0.50	-0.192	0.258	0.258	92%

Notes: Based on $R = 100$ Monte Carlo replications with sample size $N = 1{,}000$. Point estimates are posterior medians. Coverage refers to the empirical coverage of 95% HPD intervals. The model specification includes an intercept in the intensity equation but not in the participation equation.

models, the threshold parameters exhibit somewhat larger dispersion, although their coverage probabilities remain close to the nominal level. The correlation parameter ρ is also recovered reasonably well, with moderate bias and satisfactory interval coverage. These results suggest that the proposed Bayesian estimation procedure

Table B.4: Monte Carlo Results: Bayesian Zero-Inflated Ordered Probit Model (No Intercepts)

Parameter	True value	Bias	MAE	RMSE	Coverage
γ_2	-0.25	0.002	0.014	0.014	0.960
γ_3	-1.00	-0.010	0.101	0.101	0.960
β_2	1.00	0.071	0.178	0.178	0.940
β_3	-0.50	-0.028	0.103	0.103	0.960
α_1	4.50	-0.074	0.577	0.577	0.940
α_2	5.50	-0.020	0.618	0.618	0.980
ρ	0.50	-0.204	0.289	0.289	0.890

Notes: Based on $R = 100$ Monte Carlo replications with sample size $N = 1{,}000$. Point estimates are posterior medians. Coverage refers to the empirical coverage of 95% HPD intervals. The model specification does not include intercepts.

provides reliable inference for the parameters of the ZIOP model and accurately reflects the underlying uncertainty.

We also investigate the identification implications of including intercept terms in both equations of the ZIOP model. Specifically, we compare three specifications: (i) intercept in the participation equation only, (ii) intercepts in both equations, and (iii) intercept in the intensity equation only.

Tables B.2–B.4 report additional Monte Carlo experiments under alternative intercept specifications. The results highlight the importance of intercept placement for identification and estimation. Including intercepts in both the participation and ordered equations leads to weak identification and poor finite-sample performance. Omitting intercepts from both equations improves identification, although at the cost of a larger bias in the correlation parameter. Among the specifications considered, including an intercept only in the participation equation provides the most reliable estimates, with the smallest bias and coverage probabilities closest to their nominal levels.

C Definition of variables and data statistics

Table C.1: Definition of variables

Variable	Definition
Strata	0 if strata is low, 1 if it is medium and 2 if it is high
Age	Age
Male	1 for male and 0 for female
Years of education	Years of education
Married	1 if married and 0 otherwise
Children	1 if the respondent has children and 0 otherwise
Young man	1 if the respondent is a man of 25 years old or younger
Capital	1 if the respondent resides in one of the main capital cities in Colombia and 0 otherwise
Alcohol/Tobacco	1 if the respondent has consumed alcohol or tobacco before and 0 otherwise
Alcohol	1 if the respondent has consumed alcohol before and 0 otherwise
Curiosity of consumption	1 if the respondent has had curiosity of using marijuana
Access to marijuana	1 if the respondent has access to marijuana
Marijuana price	Price of a marijuana cigarette in USD

Notes: Variables obtained from the ENCSP 2019 regarding to marijuana consumption and sociodemographic characteristics. Due to most of the people in the survey reported no consumption, prices for these individuals are not observable. Therefore, we impute the data by applying the nearest neighbor algorithm as follows: first we leave out prices below the 10th percentile and above the 90th percentile of the price distribution in order to avoid bias in the imputed distribution due to extreme values; second, we impute the price of marijuana units for those individuals who report no marijuana consumption using the average price of the individuals in the same municipality and socioeconomic strata; finally, if there is not a marijuana consumer in the same municipality and the same strata, we impute the price using the average of the municipality, then the average of the strata, and ultimately, the unconditional average if there is not any match. The price of marijuana is computed using the average exchange rate in 2019 (COP/USD 3,274).

Based on the data statistics we can define the representative individual. Specifically, the representative individual is a 40-year-old woman with 12 years of education (high school) who is not married, but have children, lives in a small city or town, who has previously used alcohol/tobacco, has no curiosity of using marijuana, but who has access to it.

Table C.2: Descriptive Statistics

Variable	**Mean**	**SD**	**Min**	**Max**
Strata				
Low	0.64	0.48	0	1
Medium	0.32	0.47	0	1
High	0.04	0.18	0	1
Male	0.42	0.49	0	1
Age	39.88	13.83	18	65
Years of education	12.03	4.28	0	24
Married	0.45	0.50	0	1
Children	0.70	0.46	0	1
Young man	0.19	0.39	0	1
Capital	0.27	0.44	0	1
Alcohol/Tobacco use	0.89	0.31	0	1
Alcohol use	0.88	0.32	0	1
Curiosity of consumption	0.16	0.37	0	1
Access to marijuana	0.62	0.48	0	1
Marijuana price	0.83	0.44	0.15	3.05

Notes: Statistics obtained from the ENCSPA 2019 regarding to marijuana consumption and sociodemographic characteristics.

D Interpretation of Marginal Effects, ZIOP Coefficients and Model Fit

D.1 Marijuana Application

D.1.1 Participation

The results indicate that previous alcohol or tobacco use increases the probability of participating in the marijuana market, providing evidence consistent with the gateway hypothesis. For the representative individual defined in the Appendix, previous alcohol or tobacco use reduces the probability of non-participation by approximately 2.682 percentage points.

Access to marijuana and curiosity about marijuana use operate in the same direction. Both variables reduce the probability of non-participation, although their magnitudes differ substantially. Access to marijuana reduces non-participation by approximately 1.937 percentage points, whereas curiosity reduces it by about 50.284 percentage points. This suggests that attitudes and perceptions toward marijuana are much more strongly associated with participation than access alone.

With respect to socioeconomic characteristics, participation is higher among men, younger individuals, individuals with fewer years of education, and those living in higher socioeconomic strata in major Colombian cities. The posterior evidence is strongest for age, gender, years of education, previous alcohol or tobacco use, access to marijuana, curiosity, and the young-man indicator. In contrast, the evidence for marital status and having children is weaker.

D.1.2 Intensity

Table 2 shows that higher prices are associated with lower frequencies of marijuana consumption. In particular, a one-dollar increase in the price of marijuana reduces the probability of high-frequency consumption, $P(y = 3)$, by approximately 0.008 percentage points, the probability of moderate consumption, $P(y = 2)$, by 0.007 percentage points, and the probability of low-frequency consumption, $P(y = 1)$, by 0.151 percentage points. Conversely, the probability of no consumption, $P(y = 0)$, increases by approximately 0.166 percentage points.

These results are consistent with a highly inelastic demand for marijuana in Colombia. Although a one-dollar increase represents more than a 100% increase relative to the average observed price (see Table C.2), the resulting changes in the probabilities of consumption across all frequency categories are economically small.

With respect to the sociodemographic characteristics, the results are broadly consistent with our *a priori* expectations. The probability of belonging to a higher-frequency category of marijuana consumption is greater among men, individuals residing in higher socioeconomic strata in major cities, those with fewer years of education, unmarried individuals, and those without children. In contrast, older individuals are less likely to belong to higher-frequency consumption categories.

Finally, the estimated negative correlation parameter suggests that the unobserved determinants of participation differ from those associated with intensive consumption. Conditional on observed characteristics, individuals who are more likely to enter the marijuana market due to unobserved factors are less likely to become frequent users. Conversely, unobserved factors associated with intensive marijuana use appear to be less important for the participation decision itself. This result highlights the importance of distinguishing between participation and consumption intensity, as the two decisions may be driven by different latent mechanisms.

D.2 Tobacco Application

D.2.1 Participation

The tobacco results share several similarities with those obtained for marijuana, although some important differences emerge. As in the marijuana application, prior alcohol use is strongly associated with participation in the tobacco market, providing evidence consistent with the gateway hypothesis. However, the magnitude of the effect is substantially larger. For the representative individual, prior alcohol consumption increases the probability of participating in the tobacco market by approximately 57.024 percentage points, compared with 2.682 percentage points in the marijuana application.

With respect to socioeconomic characteristics, the overall patterns are broadly consistent with those observed for marijuana consumption. The main difference concerns age. Whereas participation in the marijuana market is higher among younger individuals, participation in the tobacco market is higher among older individuals. In addition, participation is more likely among men, individuals residing in higher socioeconomic strata, and married individuals. By contrast, years of education, having children, and residence in a major city do not appear to be important determinants of participation. These results suggest that, although both substances share some common drivers, the demographic profile of tobacco users differs from that of marijuana users, reflecting the distinct social and behavioral dynamics of the two markets.

D.2.2 Intensity

On the one hand, with respect to the sociodemographic regressors, the results indicate that the probability of belonging to a higher-frequency category of tobacco consumption is greater among men, older individuals, those residing in higher socioeconomic strata in major cities, unmarried individuals, those with children, and those with fewer years of education.

On the other hand, the estimated correlation parameter is positive, suggesting that the unobserved factors associated with participation in the tobacco market are also positively associated with intensive tobacco consumption. In other words, conditional on the observed covariates, individuals who are more likely to participate in the tobacco market because of unobserved characteristics also tend to be more likely to consume tobacco at higher frequencies. This result contrasts with the marijuana application, where the estimated correlation is negative, and suggests that participation and consumption intensity in the tobacco market are driven by a more similar set of latent factors.

Finally, recall that this interpretation is based on the marginal effects that observable covariates have onto the outcome probabilities. These are non-linear transformations of the ZIOP model parameters as shown in equations 4–6. The full estimated coefficients both the marijuana and tobacco application can be found in Table D.1.

E Convergence diagnostics

Several diagnostic measures were computed to assess the convergence and stationarity of the posterior samples. The results, reported in Table E.1, provide substantial evidence of satisfactory convergence. The Heidelberger–Welch test (Heidelberger and Welch, 1983) indicates stationarity for 24 of the 27 parameters. Likewise, the Geweke test (Geweke, 1992), which compares the posterior means from the first 10% and the last 50% of the Markov chain, is passed by 21 of the 27 parameters. Taken together, these diagnostics suggest posterior samples are reliable for inference.

Table D.1: Model results

Variable	Marijuana				Tobacco			
	ML		Bayesian		ML		Bayesian	
	Coef	SE	Coef	SD	Coef	SE	Coef	SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Participation								
Constant	-1.408	0.305*	-1.418	0.329*	-4.026	0.305*	-4.133	0.309*
Age	-0.553	0.064*	-0.602	0.157*	0.694	0.076*	0.721	0.077*
Strata								
High	0.451	0.114*	0.439	0.117*	0.331	0.113*	0.340	0.115*
Medium	0.140	0.051*	0.154	0.061*	0.154	0.048*	0.155	0.048*
Years of education	-0.023	0.006*	-0.023	0.006*	-0.002	0.005	-0.001	0.005
Male	0.189	0.051*	0.216	0.088*	0.527	0.043*	0.531	0.043*
Children	0.125	0.059*	0.098	0.091	0.087	0.050	0.087	0.051
Married	0.031	0.060	-0.012	0.120	0.139	0.049*	0.138	0.049*
Capital	0.122	0.046*	0.124	0.044*	0.025	0.043	0.030	0.044
Young man	-0.422	0.064*	-0.385	0.112*	-0.094	0.063	-0.075	0.062
Alcohol/Tobacco use	1.246	0.157*	1.235	0.211*				
Alcohol use					1.792	0.063*	1.791	0.065*
Access marijuana	0.487	0.038*	0.544	0.156*				
Curiosity	2.085	0.053*	2.068	0.106*				
Intensity								
Age	0.159	0.040*	0.255	0.279*	0.012	0.017	0.009	0.017
Strata								
High	-0.138	0.130	-0.184	0.177	0.155	0.045*	0.150	0.046*
Medium	-0.001	0.066	-0.012	0.070	0.051	0.022*	0.049	0.022*
Years of education	0.011	0.007	0.007	0.011	-0.030	0.002*	-0.030	0.002*
Male	0.322	0.063*	0.317	0.065*	0.516	0.020*	0.513	0.020*
Children	-0.312	0.070*	-0.289	0.089	0.026*	0.024	0.026	0.025
Married	-0.313	0.070*	-0.302	0.071*	-0.209	0.020*	-0.210	0.021*
Capital	0.108	0.058	0.106	0.057	0.209	0.035*	0.208	0.035*
Young man	0.573	0.061*	0.531	0.116*	0.090	0.033*	0.091	0.033*
Marijuana price	-0.183	0.034*	-0.205	0.066*				
Threshold 1	1.384	0.071*	1.223	0.650*	1.019	0.017*	1.020	0.018*
Threshold 2	1.662	0.060*	1.809	0.390*	1.164	0.031*	1.166	0.019*
Correlation	-0.501	0.144*	-0.463	0.079*	0.365	0.130*	0.348	0.055*
Regional effects					✓		✓	

Notes: Results for the marijuana and tobacco applications of the ZIOP model with correlated disturbances by the Bayesian and frequentist approaches. Coef refers to the estimated coefficients, SE refers to standard errors (for the frequentist approach), and SD refers to standard deviations (for the Bayesian approach). Columns (1), (3), (5) and (7) show the coefficient while Columns (2), (4), (6) and (8) show the standard deviation/standard error. Columns (1), (2), (3), and (4) show the results for the marijuana application. Columns (5), (6), (7), and (8) show the results for the tobacco application. Columns (1), (2), (5), and (6) show the results for the ZIOP model by maximum likelihood. Finally, Columns (3), (4), (7), and (8) show the results of the ZIOP model using the Bayesian framework. Statistically significant variables marked with $*$ based on 95% credible intervals for Bayesian methods and the significant test for the frequentist.

Table E.1: Convergence diagnosis

Variable	**Geweke**		**Heidelberger–Welch**	
	Marijuana	Tobacco	Marijuana	Tobacco
Participation				
Constant	-2.30	**-0.87**	**Passed**	**Passed**
Age	4.28	**1.06**	**Passed**	**Passed**
Strata				
High	**-0.15**	**1.14**	**Passed**	**Passed**
Medium	**-0.68**	**1.07**	**Passed**	**Passed**
Years of education	**-0.56**	**-1.62**	**Passed**	**Passed**
Male	-4.00	**0.11**	**Passed**	**Passed**
Children	**1.16**	**-0.90**	**Passed**	**Passed**
Married	**1.61**	**-1.09**	**Passed**	Failed
Capital	**0.70**	**1.19**	**Passed**	**Passed**
Young man	**0.19**	**1.85**	**Passed**	**Passed**
Alcohol/Tobacco use	**1.32**		**Passed**	
Alcohol use		**-0.49**		**Passed**
Access marijuana	-2.71		**Passed**	
Curiosity	2.60		**Passed**	
Intensity				
Age	-5.32	**-0.02**	**Passed**	**Passed**
Strata				
High	**1.44**	**-1.66**	**Passed**	**Passed**
Medium	**1.21**	**-1.81**	**Passed**	**Passed**
Years of education	2.73	3.01	**Passed**	**Passed**
Male	2.93	**-0.16**	**Passed**	**Passed**
Children	**-0.30**	**0.81**	**Passed**	**Passed**
Married	-2.07	**0.68**	**Passed**	**Passed**
Capital	**-0.69**	**-0.73**	**Passed**	**Passed**
Young man	**0.16**	**1.95**	**Passed**	**Passed**
Marijuana price	**0.89**		**Passed**	
Threshold 1	5.40	2.04	**Passed**	**Passed**
Threshold 2	-2.94	2.03	**Passed**	**Passed**
Correlation	**-0.10**	**0.20**	**Passed**	**Passed**

Notes: Converge diagnosis for the posterior samples of the ZIOP model. We present the Geweke (Geweke, 1992) and Heidelberger-Wech (Heidelberger and Welch, 1983) tests. For the Geweke test, we show the z-score, with passing coefficients at 5% of significance. Finally, we provide the results of the Heidelberger-Welch test. Bold font indicate passing a test.