Contents

Universidad Nacional Autónoma de México

Facultad de Ciencias

Proyecto Final

Estadística Bayesiana

Castro Gómez Pedro Pablo

Fernández García Edson Jehovani

Martínez Herrera Tania Melisa

Pimentel Bolívar Luis Emmanuel

1 Introducción

La Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE) es un instrumento realizado y aplicado por el INEGI con múltiples objetivos. Entre ellos se encuentran medir la victimización personal y del hogar, estimar el número de víctimas y número de delitos ocurridos a lo largo del año y realizar algunas otras mediciones y estimaciones relevantes sobre victimización y percepción de la seguridad de los mexicanos. Todo esto se realiza con el fin de generar información disponible para el público general y que sea de utilidad para la implementación y mejora de políticas públicas en la materia. La ENVIPE arroja como resultado información en forma de tabulados con estimaciones y bases de datos de las respuestas obtenidas.

Para la realización de este proyecto, se trabajó con la base de datos de respuestas de la encuesta realizada en 2019, centrándose en estudiar y modelar el número de robos y/o asaltos ocurridos a los encuestados cuya residencia se encuentra en la CDMX. En particular, es de interés contestar a los cuestionamientos: ¿existe alguna relación entre ser víctima de robo y las condiciones y/o características de las personas (de acuerdo con los datos que recopiló la encuesta)? ¿El número de veces que se es víctima a lo largo del año puede tener relación con dichas características?

Para contestar estas preguntas, se realizó el análisis y posterior modelación tanto de la probabilidad de sufrir algún robo o asalto como del número de veces que se fue víctima a lo largo del año de alguno de estos delitos. Para ambas variables se consideraron diversas características de los encuestados.

2 Limpieza de la Base

En un principio, se tomó a consideración aquellas preguntas realizadas que a primera instancia pareciesen influir en el valor de nuestras variables respuesta. Así, se consideró como opción a las siguientes covariables:

1. Sexo
2. Edad
3. Nivel educativo (Niv_Edu)
4. Nombre de la alcaldía (Nom_Mun)
5. Situación laboral (Sit_Lab_Act)
6. Posición ocupacional (Pos_Ocup)
7. Importancia de la seguridad para la persona en su localidad (Imp_Seg)
8. Considera segura su localidad (Seg_Loc)
9. Considera segura su alcaldía (Seg_Mun)
10. Problemas en su localidad de: Alumbrado (Alum), agua (Agua), pandillas (Pandill), robos (Robos) o delincuencia en las escuelas (Del_Esc)
11. Aumento de operativos contra la delincuencia en su localidad (Mas_Op_Del)
12. Aumento de patrullas de vigilancia en su localidad (Mas_Pat_Vil)
13. Alguien de su hogar posea un vehículo (Vehic).

Las primeras anotaciones con respecto a estas candidatas fueron el exceso de respuestas a varias de las preguntas cuya connotación fuera no útil para la creación de un modelo, ya que dichas respuestas eran análogas a un “No sé”; en segunda instancia, lo que se puede analizar en el siguiente correlograma:

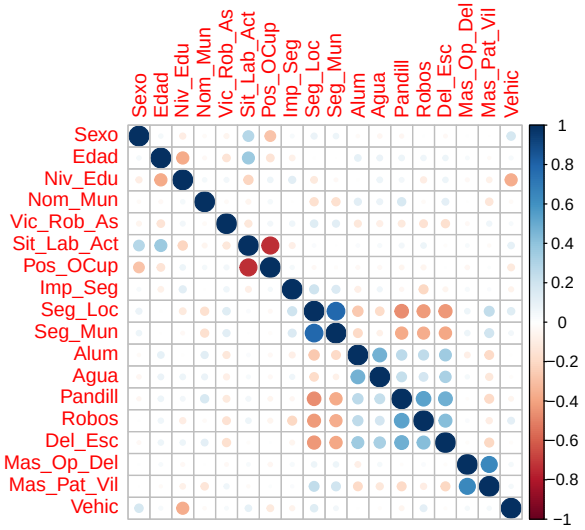


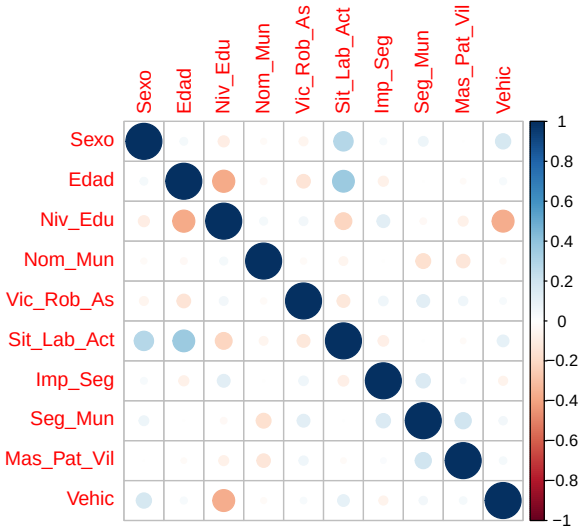
Figure 1: Correlograma Variables Iniciales

Se puede apreciar una fuerte correlación entre las variables Seg_Loc, Seg_Mun, Alum, Agua, Pandill, Robos y Del_Esc. Esto tiene sentido pues, además de que provienen de la misma sección de la encuesta, se puede suponer por las últimas el tipo de localidad en la que vive el encuestado, y a partir de esto, es sencillo deducir cómo son las condiciones de seguridad de la zona.De esta forma, por el fuerte vínculo entre las covariables es por lo que se decidió quedar solamente con la percepción de la seguridad en su alcaldía, pues al final es por esta categoría que se tienen identificados en otra covariable.

Otras covariables que cuentan con un fuerte vínculo son Pos_Ocup y Sit_Lab_Act. Esto es debido a que en las respuestas en situación laboral se encuentran divididas de tal forma que se identifiquen a aquellos que trabajen dentro de las primeras opciones y a los que no en las posteriores; en el caso de Pos_Ocup, las primeras opciones están relacionadas a que trabajen y la última a que no lo hacen. En este caso, se decidió sólo considerar la covariable de Sit_Lab_Act, pues así se tiene una mejor noción del tipo de responsabilidades que tiene la persona y la exposición que tienen a ser asaltados o robados.

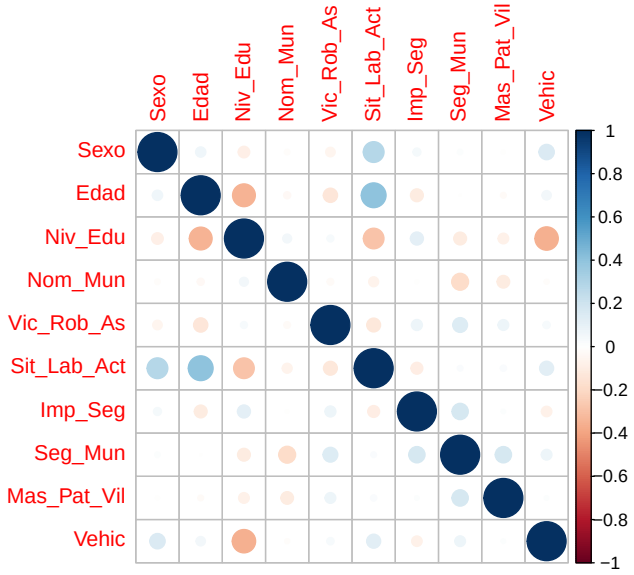
Existen otras covariables que parecen poseer una correlación significativa, pero en esos casos no se logró dar una razón clara por la cual sólo elegir una de ellas, de tal forma que conservamos el resto.

Una vez realizada la nueva selección de covariables, se obtiene el siguiente correlograma.



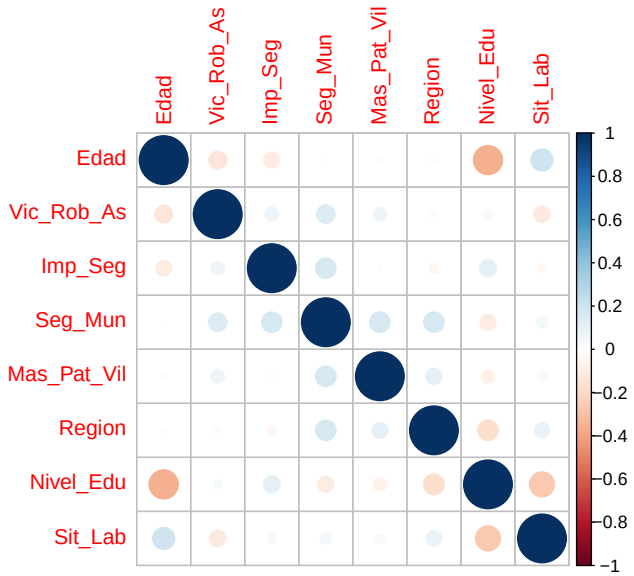
El siguiente tema a abordar es la imputación de datos. Esto es necesario, pues si se eliminaran todas las observaciones en que se respondió de forma "desconocida" o no aplicable para un modelo, se estaría eliminando aproximadamente el 20% de las observaciones y con ello causando un sesgo importante. La imputación realizada consiste en sustituir las respuestas no informativas para el estudio por aquéllas que sí lo sean tomando en cuenta la proporción de cada respuesta útil con respecto a nuestra variable respuesta y de esta forma evitar un sesgo en cuanto a la correlación con respecto a ésta y con el resto de covariables.

Una vez realizada la imputación, se hace presente este correlograma:



Se observó que efectivamente no se ha afectado de manera notoria a la relación que tenían las covariables entre sí y de éstas con la variable respuesta.

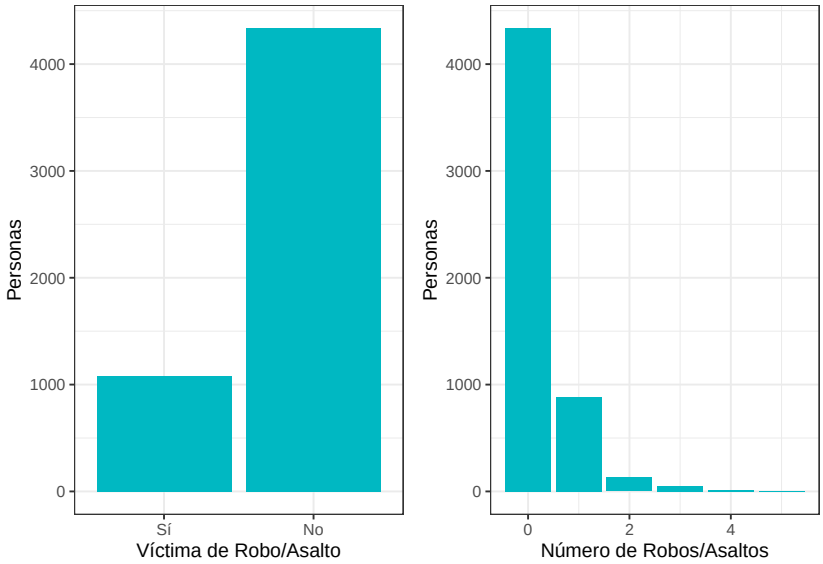
A la hora de la realización de los modelos ajustados, se presentan algunas covariables que sería útil categorizar de otra forma para una mejor interpretación. Dichas covariables son: Nom_Mun, Niv_Edu, Sit_Lab_Act. A su vez, se decidió eliminar las variables "sexo" y "vehic" pues no presentaban relevancia para la variable respuesta. Finalmente, llegamos a nuestro último correlograma, que se muestra a continuación.



3 Análisis Descriptivo

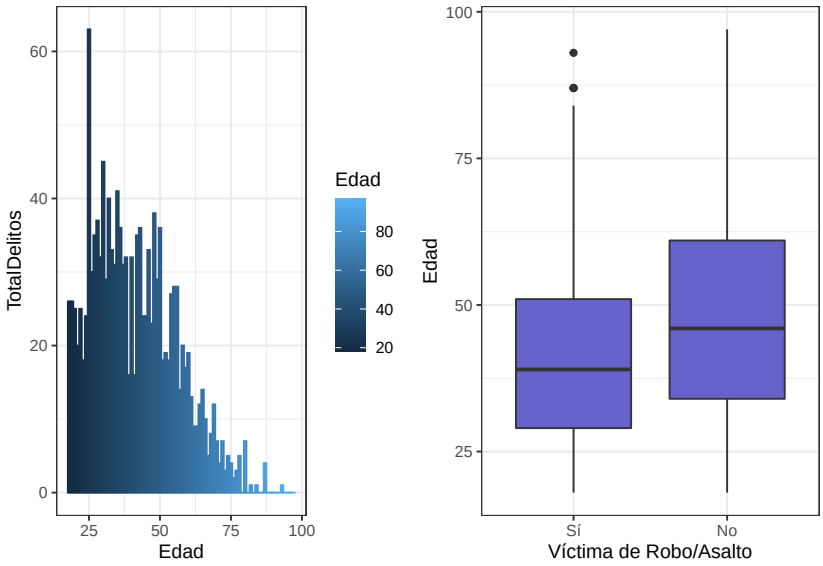
En esta sección se analizarán las variables que se conservaron, lo que ayudará a conocer mejor la base.

Se tienen dos variables respuestas. La primera es la cantidad de personas que han sido robadas o asaltadas, mientras que la segunda es la cantidad de veces que las personas fueron víctimas de estos delitos. La composición de ambas variables puede ser vista en las siguientes gráficas:



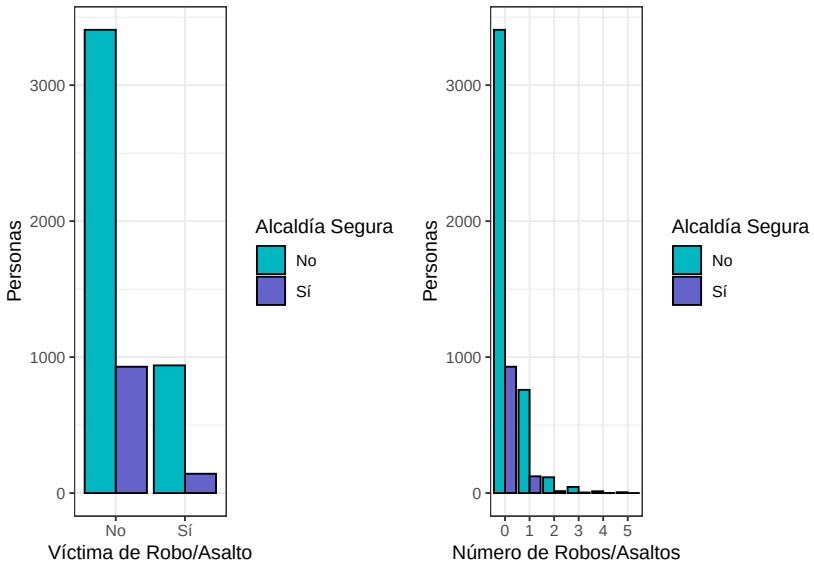
Se puede destacar que en ambas variables respuesta, el número de observaciones relacionadas a la presencia nula de siniestros tiene una proporción mayor a aquéllas donde la persona fue víctima de robos y/o asaltos. Igualmente, se observa que una vez que condicionamos a que la persona haya sufrido un altercado, hay una probabilidad muy alta de que sólo haya sido una vez a que haya sido múltiples veces.

La primera covariable a analizar es la edad.



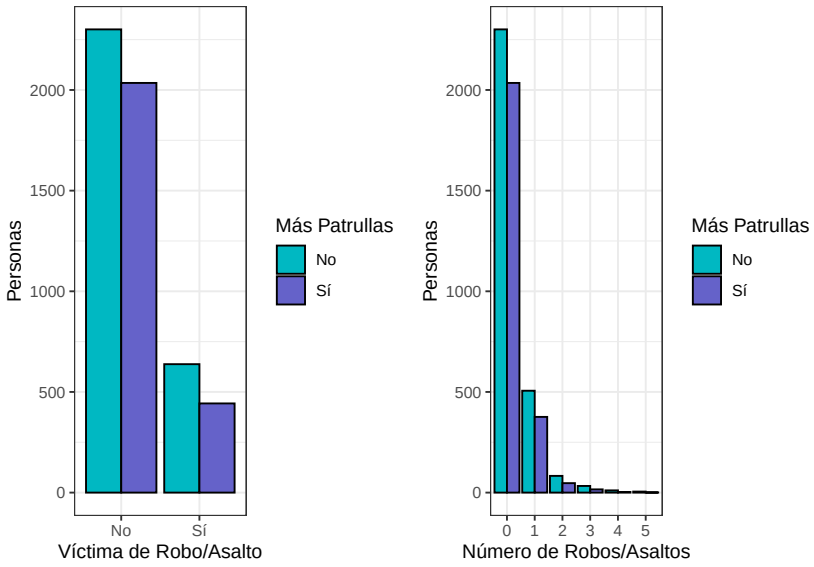
Se notó es que, a medida que se llega a edades donde la persona probablemente se encuentre jubilada, retirada o en condiciones no aptas para seguir trabajando, hay una disminución en la cantidad de robos y asaltos que esta sufre. En cambio, en edades en que la persona se encuentra dentro del sector de la población económicamente activa, se sigue una tendencia relativamente uniforme de siniestros.

Como segunda covariable se tiene la respuesta a la pregunta de si la persona considera segura su alcaldía o no.



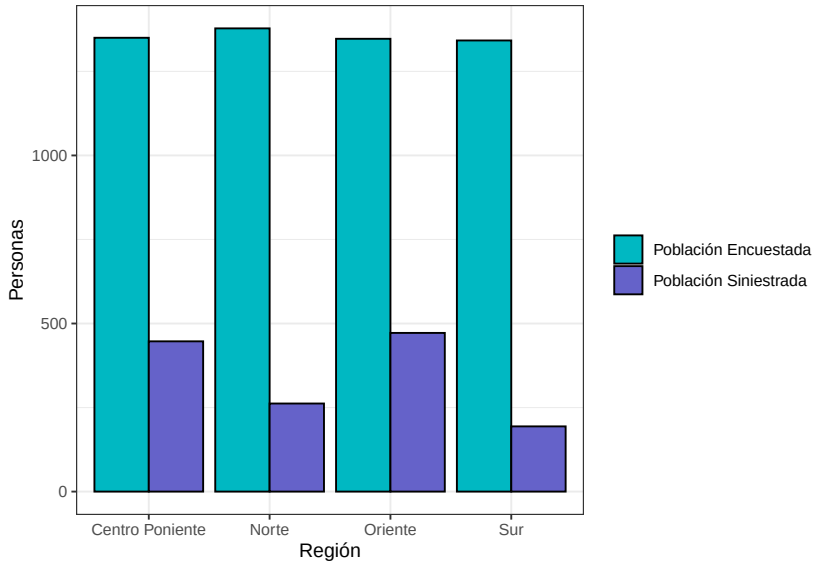
Como era de esperarse, se nota el efecto de que consideren su alcaldía más segura aquellas personas que no han sufrido algún altercado.

Como tercera covariable se tiene la respuesta a la pregunta de si la persona percibió un intento por parte de su gobierno local por incrementar la cantidad de patrullas de vigilacia en la zona.



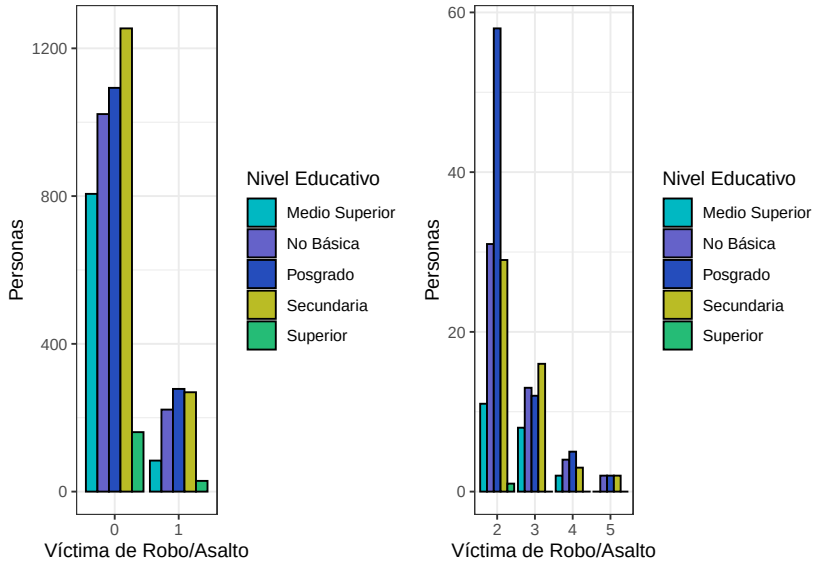
A pesar de no notar un cambio evidente en la proporción de asaltos en aquellas localidades donde se observó un aumento en las patrullas de vigilancia, al momento de incluirla como covariable en los modelos se notó que sí tenía relevancia. Esta es la razón por la que se conservó esta covariable.

Inicialmente se contaba con la covariable de la alcaldía en donde se residía, sin embargo, al tenerla de esta forma, la influencia de esta covariable en los modelos no era significativa en lo absoluto. Como consecuencia, se decidió agrupar a las alcaldías como lo realiza la misma ENVIPE, que es por zona geográfica.



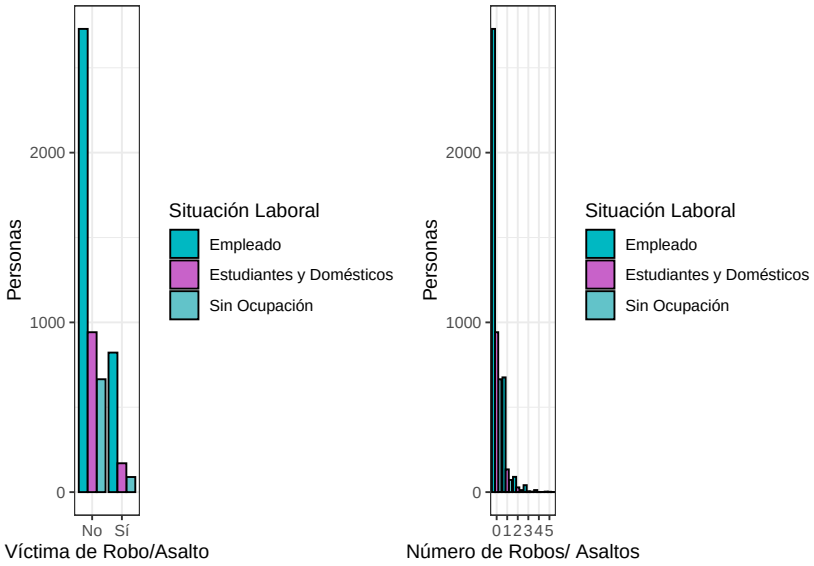
En las gráficas es posible ver que los habitantes de la región Centro Poniente y Oriente suelen estar más afectados a comparación de las regiones Norte y Sur.

Como siguiente covariable, se considera el máximo nivel educativo en el cual se desempeñó el encuestado.



Se puede observar que los que han estudiado un posgrado suelen ser asaltados más veces en comparación. Por su parte, los encuestados que hayan llegado a niveles No-Básico y Secundaria suelen mantenerse similares.

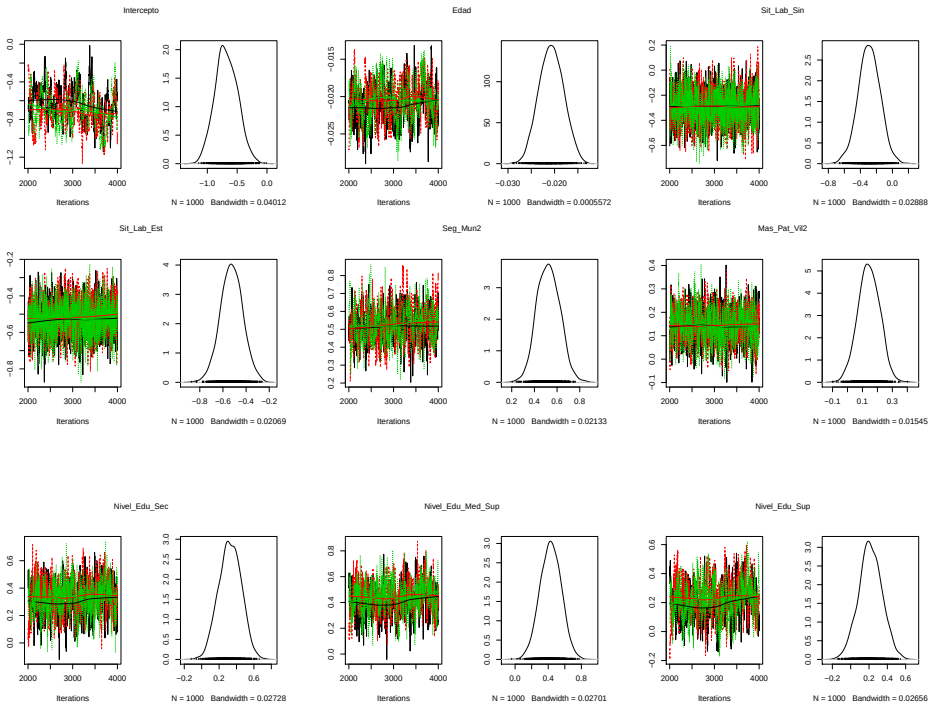
En la covariable Situación Laboral Actual, se agruparon de esta forma las categorías al suponerse que los trabajadores serían los más afectados. Al contrario de los de Sin Ocupación, que presentarían menor riesgo. Mientras que los Estudiantes y Domésticos tendrían un riesgo medio porque tienen que salir constantemente a la calle, pero no suelen llevar tanto dinero consigo.

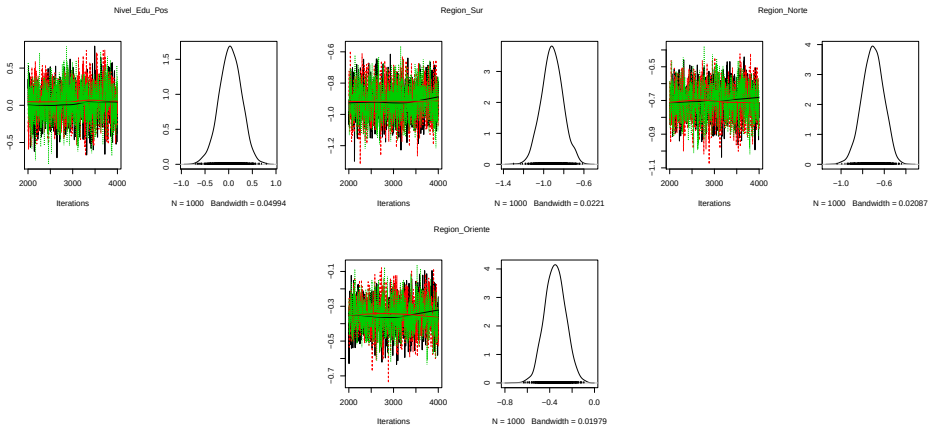


4 Modelo Logit

El modelo logístico es Se decidió implementar un modelo logit para estimar únicamente la probabilidad de que a una persona la asalten.

Modelo	AIC	Punto de Corte	Efectividad Clásica	Efectividad Simulada
Modelo Completo	5107.476	0.2083094	62.51	77.15
Modelo 1	5108.358	0.2076318	62.06	77.20
Modelo 2	5114.248	0.2087943	62.54	77.02





Variables	Clásico	Bayesiano
Intercepto	-0.490	-0.680
Edad	-0.020	-0.021
Sit_Lab_Sin-Oc	-0.291	-0.290
Sit_Lab_Est-Dom	-0.526	-0.525
Seg_Mun_2	0.507	0.524
Mas_Pat_Vil_2	0.178	0.143
Nivel_Edu_Sec	0.001	0.325
Nivel_Edu_Med-Sup	-0.342	0.429
Nivel_Edu_Sup	0.084	0.215
Nivel_Edu_Pos	0.056	0.035
Region_Sur	-0.912	-0.916
Region_Norte	-0.700	-0.706
Region_Oriente	-0.350	-0.353

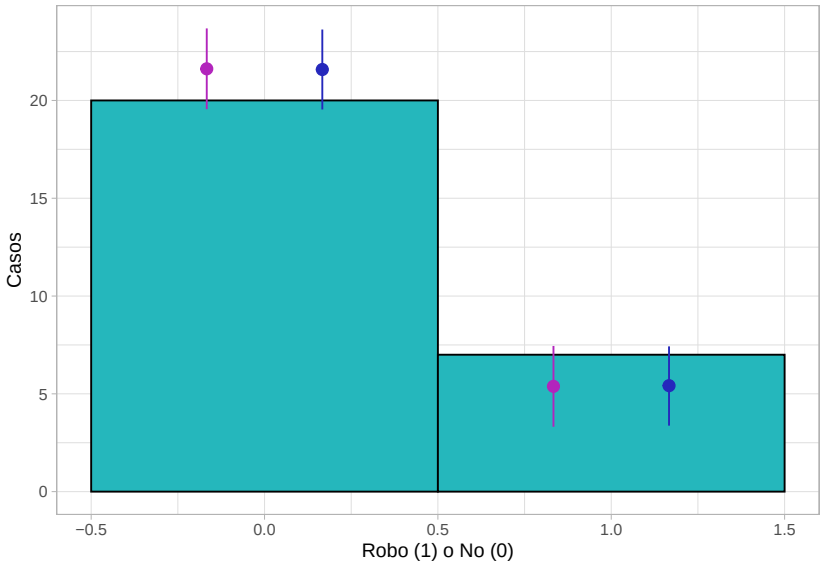
4.1 Ejemplo

A continuación, a modo de ejemplo gráfico, se quiso considerar a todas las personas con las siguientes características específicas: personas de 22 años que sean estudiantes de licenciatura o que ya la hayan concluido, pero sin haber iniciado cursos posteriores.

De esta forma, consideramos todas las posibles combinaciones de las demás covariables no especificadas y obtuvimos los valores ajustados de las probabilidades tanto para el modelo clásico como para el bayesiano. De igual manera, realizamos simulaciones con las probabilidades estimadas de ambos modelos considerando las observaciones reales en cada subconjunto y se decidió comparar con los datos reales de la encuesta. Así, llegamos al siguiente resultado.

Table 1: Ji-cuadrada test con p-values e interpretación

p-value	Interpretación
3.910882e-86	ZIP > modelo nulo



Se aprecia que ambos modelos contienen al valor real dentro de sus intervalos de confianza (dado por su desviación estándar), por lo que, a pesar de que la media estuvo un poco alejada del valor real, se tiene una idea de los valores que puede adquirir nuestra variable respuesta. Además, sí se notó una muy leve mejora entre lo obtenido con el modelo bayesiano (azul oscuro) y el modelo clásico (morado), pero ambos funcionan de manera muy similar, pues los valores de los coeficientes obtenidos son demasiado cercanos entre sí.

5 Modelo Poisson Cero Inflado

PARTE DE PABLO

Parte de las pruebas

Para comprobar que es un buen modelo, se realizaron ciertos **tests** y se tomaron en cuenta algunos criterios. Entre ellos, la prueba de la Ji-Cuadrada utilizando las verosimilitudes de un modelo poisson nulo (es decir, solo se toma el intercepto como variable predictiva) contra la de un modelo poisson inflado. Esto ayuda a ver que evidentemente el modelo tiene un mejor ajuste a los datos que si solo se considerara un modelo con ninguna otra covariable (cita).

Por otro lado, también se consideró el Vuong test, ya que este indica si el modelo Poisson Inflado se ajusta mejor que solo considerar una regresión Poisson para el desarrollo del mismo (cita).

OJO: Creo que aquí hay que anexar la parte de porqué es mejor ajustar un modelo Binomial Negativo, la respuesta es porque no tenemos mucha sobredispersión, ya que la varianza de nuestros datos (0.3449009) no es mucho más grande que la media de los mismos (0.2538305) (Cita). Pero, si no les convence tanto, aquí hay una página : https://support.sas.com/resources/papers/sgf2008/countreg.pdf

Dichas pruebas se resumen en las siguiente tablas, en donde se muestran los respectivos **p-value** y su interpretación.

Table 2: Vuong test con p-values e interpretación

	p-value	Interpretación
Raw	<2.22e-16	ZIP > modelo poisson
AIC-corrected	2.8866e-15	ZIP > modelo poisson
BIC-corrected	3.0909e-11	ZIP > modelo poisson

Table 3: Coeficientes estimados del modelo poisson inflado para la parte Logit utilizando JAGS vs el enofoque clásico

Variables	Bayes	Clásico
Intercepto	3.6945	-3.3901
Edad	-0.0434	0.0411
Mas_Pat_Vil_2	0.4899	-0.4456
Region_Sur	-1.1479	1.2300
Region_Norte	-1.3946	1.2720
Region_Oriente	-2.2834	2.0829
Sit_Lab_Sin_Oc	-0.5789	0.5504
Sit_Lab_Sin_Dom	-0.8706	0.8450

Se puede apreciar que tanto la prueba Ji como el Vuong Test, dieron resultados a favor del modelo ajustado. Es decir, este es estadísticamente significativo mejor que considerar solo el modelo Poisson o utilizar un modelo nulo. De esta manera, se puede corroborar que tiene sentido ajustar un modelo poisson cero inflado.

De igual forma, realizamos una comparación mediante una simulación con los valores reales de la base para las covariables, realizada tanto con el modelo cero inflado como con un modelo de regresión Poisson simple. Obteniendo el resultado observado en la siguiente gráfica.

Gráfica

Mañana me avisas, Pablo, si esta interpretación de la gráfica me tocaba a mí. Es que tú fuiste quien la explicó en la expo, y no sé si ya tengas algo planeado para esta parte.

5.1 Estimación de parámetros desde el enfoque bayesiano

Ahora, considerando un ajuste desde el enfoque Bayesiano, se obtuvo la estimación de los siguientes parámetros de un modelo Poisson compuesto cero inflado.

5.2 JAGS

Se puede observar que en general, las estimaciones son bastantes similares al enfoque frecuentista. No obstante, los signos para la parte relacionada al modelo logit considera los opuestos al modelo clásico, esto se debe a la implementación para calcular dichos estimadores tanto en el frecuentista como en el bayesiano; mientras uno se basa en sacar la probabilidad de cero, el otro se basa en sacar la de 1. Sin embargo, en la literatura esto se ignora para el caso de la interpretación de los estimadores (e.g., Atkins & Gallop, 2007; Ravert, Schwartz, Zamboanga, Kim, Weisskirch & Bersamin, 2009; Lewis et al., 2010). Por lo que se basó en la interpretación del modelo frecuentista.

Por otro lado, es importante ver que nuestros parámetros estimados desde enfoque bayesiano tengan estimaciones adecuadas. Para lograr esto, se puede acudir a la visualización de gráficas que indiquen si las cadenas

Table 4: Coeficientes estimados del modelo poisson inflado para la parte poisson utilizando JAGS vs el enofoque clásico

Variables	Bayes	Clásico
Intercepto	-1.3013	-1.2615
Seg_Mun_2	0.4780	0.4616
Region_Sur	-0.4538	-0.3927
Region_Norte	-0.1911	-0.1870
Region_Oriente	0.7526	0.7417

de markov subyacentes convergen a las distribuciones posteriores de los parámetros. A continuación, se muestran dichas gráficas.

Nota: En el caso presentado, se utilizaron dos cadenas distintas, ya que el rendimiento computacional era mucho más grande a partir de la tercer cadena.

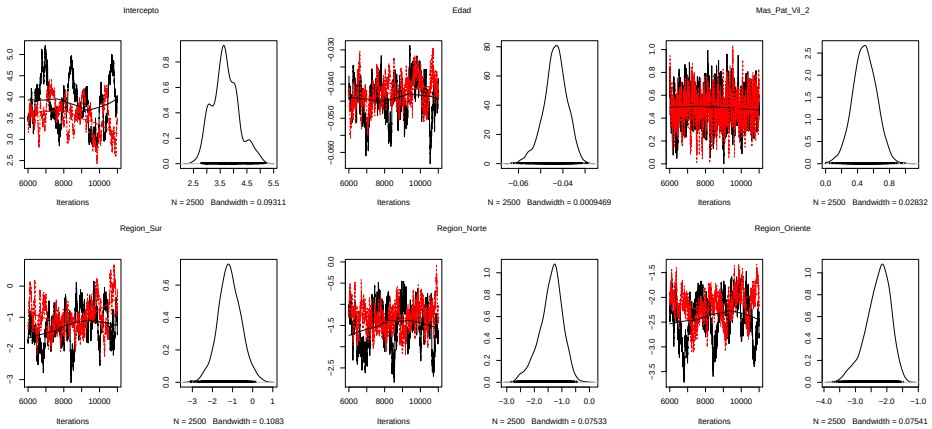


Figure 2: Distribuciones posteriores de parámetros

A través del software estadístico de JAGS, se pudó lograr obtener las distribuciones posteriores de los parámetros para el modelo Poisson Inflado; y vemos que efectivamente en casi en todos los casos se llega a la convergencia. Además, se verifico que la mayoría de nuestros estimadores fueran significativos, por lo cual podemos tener parámetros bien ajustado.

5.3 Diferencias entre el enofoque Bayesiano y el clásico

Para esta parte, se generaron simulaciones del modelo bayesiano para ver que tan bien se ajustaban estos datos a los reales y se obtuvieron los siguientes resultados:

Gráfica modelo Bayesiano vs Frecuentista

Se puede apreciar, por un lado, que existe una mejoría en cuanto a la eficiencia del modelo por .4%. Es decir, en las simulaciones coincidido un mayor porcentaje con respecto a los datos reales. Sin embargo, también se puede ver que el modelo frecuentista gana en acercarse a la media de los datos y se ajusta mejor para las primeras categorías del histograma.

En resumen, a pesar de las diferencias entre un modelo y otro, es importante reconocer el buen ajuste del modelo bayesiano y su proximidad al modelo frecuentista. Además, estamos considerando distribuciones aprioris no informativas acerca de los datos, y ya que el enfoque bayesiano otorga una mayor flexibilidad al recabar más información, se pueden inducir nuevas estimaciones que se puedan ajustar mejor a los datos.

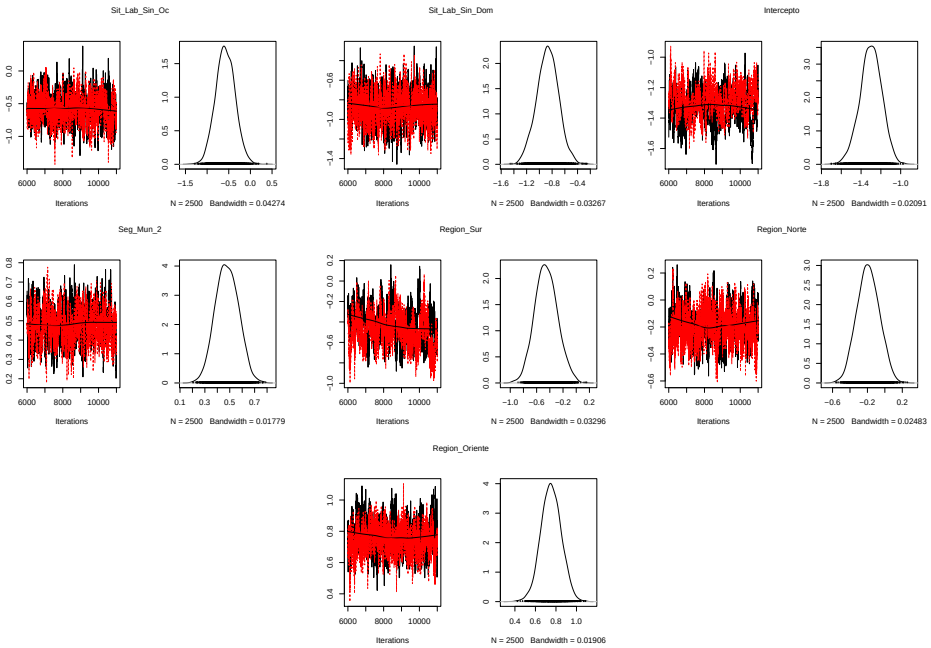


Figure 3: Distribuciones posteriores de parámetros

También, es importante destacar que existe poca literatura que trata el caso específico del modelo Poisson Inflado y este asunto se agrava aún más si se intenta implementar mediante el enfoque bayesiano.

6 Conclusiones

En conclusión, para el caso de nuestro estudio, se tienen dos modelos que pueden ayudar a modelar el comportamiento de estos delitos y, de igual manera, de los factores que puedan influir en la ocurrencia de los mismos.

Esta información puede apoyar al desarrollo de políticas públicas y acciones que mejoren la seguridad. Por ejemplo, dar mayor atención a aquellas áreas propensas a este tipo de crímenes, ver si realmente existe alguna relación entre el número de patrullas y la percepción de la seguridad o, simplemente, identificar a la población más vulnerable. Así, poder tener más herramientas para entender este fenómeno y atenderlo de una manera informada y oportuna.