MODELO DE REGRESIÓN DE POISSON INFLACIONADO EN CEROS, EN EL MODELAMIENTO DE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD DEL QUINDÍO.

Trabajo para optar al Título de:
Magister en Enseñanza de las Matemáticas.

AUTOR: Andrés Mauricio Ramírez Gonzaléz.
Código: 9736830

Directora: Ph.D. Diana Milena Galvis Soto

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE CIENCIAS BÁSICAS
MAESTRÍA EN ENSEÑANZA DE LA MATEMÁTICAS
2017

MODELO DE REGRESIÓN DE POISSON INFLACIONADO EN CEROS, EN EL MODELAMIENTO DE LA DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD DEL QUINDÍO.

AUTOR: Andrés Mauricio Ramírez Gonzaléz. Código: 9736830

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE CIENCIAS BÁSICAS
MAESTRÍA EN ENSEÑANZA DE LA MATEMÁTICAS
2017

Nota de aceptación

Firma de Director de Tesis

Firma de Jurado de Tesis

Firma de Jurado de Tesis

Agradecimientos

- A Dios que me brinda la fortaleza necesaria para enfrentar cada día.
- A mi directora de trabajo Diana Milena Galvis Soto por su tiempo, paciencia, claridad,asesoría continua, orientación y aportes permanentes.
- A la profesora Gladys Elena Salcedo Echeverry por la oportunidad y la confianza.
- Al Grupo de Investigación y Asesoría en Estadística de la Universidad del Quindío por la disposición y el apoyo brindado para la realización de este trabajo.
- A mi grupo de trabajo en la maestría Ana, Anghela y Jhon Faber por el apoyo incondicional en las largas jornadas de estudio que llevaron a crear lazos de amistad.
- A los docentes y directivos de la Universidad Tecnológica de Pereira por sus enseñanzas y acompañamiento.
- A cada una de esas personas que me han apoyado, animado y orientado en este largo proceso.

Índice general

Índice de figuras

Índice de tablas

Resumen

Datos obtenidos mediante conteo surgen en diversas áreas del conocimiento y pueden ser modelados mediante distribuciones discretas como la Poisson o la Binomial Negativa; sin embargo, una característica común en algunas situaciones experimentales es contar con muestras que contienen una cantidad de ceros que excede la esperada para estos modelos. En este caso, se dice que el conjunto de datos presenta un exceso de ceros y por lo tanto para realizar su análisis Lambert (1992) propone el modelo de regresión de Poisson con exceso de ceros (ZIP por sus siglas en inglés).

En este trabajo se propone emplear este modelo como una herramienta que permita identificar con alto valor de probabilidad, a aquellos estudiantes que se encuentren en riesgo de convertirse en desertores tempranos de alguno de los programas académicos ofertados por la Universidad del Quindío, y con ello, brindar información necesaria y efectiva a las dependencias correspondientes que permita a la Institución Educativa ajustar estrategias de intervención que ayuden a disminuir los niveles de deserción lo que conlleva a un aumento en la tasa de graduación.

En la aplicación del modelo se utilizaron los datos de 288 estudiantes de la Universidad del Quindío que ingresaron a primer semestre durante el año 2012 y que abandonaron sus estudios por diversas causas y en diferentes momentos de su formación académica. En este caso la variable respuesta considerada es el número

total de materias aprobadas por cada estudiante hasta el momento de abandonar sus estudios; debido a que muchos de ellos toman esta decisión al inicio del primer semestre, se observa una cantidad de ceros mayor a la esperada.

Con base en esta información fue ajustado el modelo ZIP y una vez obtenidas las estimaciones de los parámetros y junto con los valores de las variables observadas se calcula para cada estudiante la probabilidad de abandonar sus estudios con cero, una, dos, ... y hasta diez materias aprobadas. Una vez obtenidas estas probabilidades se identifica a aquellos posibles desertores tempranos a través de la máxima probabilidad obtenida.

Capítulo 1

Introducción

1.1. Planteamiento del Problema

El abandono de un programa académico a nivel universitario, impacta directamente a toda la comunidad. En principio, el individuo ve trastornado su desarrollo académico lo que posteriormente conlleva a la disminución de oportunidades para ingresar a un mercado laboral el cual es cada vez más competitivo, puesto que no podrá desenvolverse de acuerdo a las exigencias necesarias por falta de conocimientos específicos. Luego están los núcleos familiares, los cuales observan con preocupación, como las posibilidades de inclusión de sus seres queridos en el grupo de profesionales que lideran la construcción de un nuevo país se aleja, igual que las posibilidades de mejorar su calidad de vida, posteriormente, las Instituciones de Educación Superior (IES) se ven afectadas, en la forma como son percibidas por la comunidad educativa; la retención de sus estudiantes y su posterior graduación dan fe del proceso exitoso de cada uno de sus programas y por otra parte se ve afectada la sostenibilidad de los ingresos percibidos con los cuales se garantizan el desarrollo de las actividades institucionales. Finalmente, se encuentra el Estado, el cual hace una inversión en infraestructura, logística, capacitación y demás, que no se ve reflejada en el desarrollo y la evolución de las comunidades cuando los

estudiantes desertan del sistema educativo.

En el trabajo *"La deserción estudiantil en la Universidad del Quindío, diagnóstico y estrategia de intervención"* realizado por el grupo de Investigación y Asesoría en Estadística de la Universidad del Quindío (Galvis et al., 2010), se puede evidenciar como en los dos primeros semestres se presenta un gran porcentaje de la deserción estudiantil, y afirman que aproximadamente un 50 % de los estudiantes que ingresan al sistema de educación superior, logra obtener su título profesional. Más detalladamente, se tiene que alrededor de un 22 % de los estudiantes que ingresan, se retiran al finalizar el primer semestre y al concluir el segundo semestre, se ha retirado aproximadamente un 30 % de estos estudiantes, lo que equivale a un poco más de la mitad de la deserción total por cohorte. Esta situación es la que justamente motiva este trabajo, pues si se logra disminuir esta deserción llamada deserción temprana se tendría un aumento en la tasa de graduación.

1.2. Justificación

Desde hace ya algún tiempo, las directrices establecidas por parte del gobierno nacional buscan cerrar las brechas en acceso y calidad a la educación superior en todas y cada una de las regiones del territorio nacional, con el objetivo de hacer de "***Colombia la más educada en el 2025***", y lograr obtener unos altos niveles educativos que promuevan la participación activa de los integrantes en el proceso de enseñanza-aprendizaje y le permitan encontrar un lugar relevante dentro de la educación latinoamericana. Se busca la construcción de una comunidad competitiva que permita el crecimiento económico de la nación y que teniendo como pilar fundamental la educación disminuya las situaciones de desigualdad social, para lo cual es imprescindible tener en cuenta las necesidades actuales de los estudiantes, de los docentes y de las IES.

Por tal motivo el gobierno nacional enmarcó sus políticas con respecto a las dificultades que presenta el sistema educativo nacional en el Plan Nacional de Desarrollo 2014-2018 "***Prosperidad para todos***"; con las cuales busca brindar solución a las problemáticas relacionadas con la calidad educativa, los niveles de cobertura institucional, la ampliación de cupos estudiantiles y la disminución en la deserción estudiantil en todos los niveles de tal forma que se fortalezcan las condiciones que posibilitan el desarrollo social de un país.

La información sobre los niveles deserción estudiantil universitaria son monitoreados por el Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior (SPADIES) y es presentada, entre otros, en procesos de acreditación de alta calidad de los programas académicos, y en consecuencia, es de interés en las IES implementar proyectos de apoyo educativo que les permitan aumentar sus tasas de retención estudiantil y posterior graduación.

Un primer paso para diseñar una estrategia de clasificación y control, que busque la disminución efectiva de la deserción estudiantil a nivel universitario, es identificar los factores que la determinan. Sin embargo, el sólo hecho de identificar estos factores no necesariamente se convierte en una solución frente al problema pues si por ejemplo, se identifica que las mujeres o los hombres son más propensos a abandonar sus estudios, nada puede hacer la institución frente a este hecho, sin contar que no se pudieran intervenir todos los hombres o todas las mujeres. Por eso más allá de la identificación de factores, la propuesta en este trabajo es utilizar un modelo estadístico que permita asignar a cada estudiante que ingresa a una IES, específicamente a la Universidad del Quindío, una probabilidad de desertar tempranamente pues como fue mencionado anteriormente la deserción se presenta, principalmente, en los primeros dos semestres académicos.

1.3. Objetivos

1.3.1. General

Desarrollar un modelo estadístico que contribuya a la identificación de los estudiantes de la Universidad del Quindío que estén en riesgo de dejar sus estudios tempranamente por factores individuales, académicos y socioeconómicos.

1.3.2. Específicos

- Ajustar el modelo de regresión de Poisson con exceso de ceros usando como variable respuesta el número de materias aprobadas por cada estudiante antes de desertar.
- Construir una función de riesgo, con base en el modelo de regresión ZIP, que cuantifique la probabilidad de desertar tempranamente que tiene un estudiante.

1.4. Metodología

En este trabajo se utilizan datos obtenidos en un estudio realizado por cohorte en la Universidad del Quindío, donde se siguió uno a uno 288 estudiantes de programas de metodología presencial que ingresaron a primer semestre durante el año de 2012 y que abandonaron sus programas académicos por diferentes causas y en distintos momentos. Una vez obtenidos estos datos se procede a ajustar el modelo de regresión ZIP y a estimar sus parámetros a través de Inferencia Bayesiana; posterior a esto y con los valores de las variables de los estudiantes en la muestra se calcula probabilidad de abandonar con cero, una, dos, ... y hasta diez materias. Esta probabilidad puede considerarse una "función de riesgo" con la que se pueden identificar desertores tempranos.

Capítulo 2

Deserción Universitaria en Colombia

La deserción universitaria es una problemática que incorpora diferentes principios (individuales, familiares, económicos e institucionales) por tal motivo su estudio ha sido abordado por diferentes áreas del conocimiento (la psicología, la sociología y las ciencias económicas), las cuales inicialmente tratan de identificar los factores relacionados y posteriormente determinar estrategias de intervención efectivas que disminuyan la cantidad de estudiantes que optan por no seguir en un programa académico determinado dentro de una Institución Educativa.

Un enfoque desde la perspectiva psicológica centra todas sus apreciaciones en el estudiante, en donde los rasgos de personalidad son factores determinantes y diferenciales, entre los estudiantes que culminan exitosamente su proceso académico y aquellos que toman la decisión de no proseguir con su formación académica, es así como Ajzen and Fishbein (1977) afirman que la decisión de no continuar con el proceso de formación está influido de forma directa por el debilitamiento de las intenciones iniciales de los estudiantes así como de su percepción de la vida universitaria. Attinasi (1986) establece como el cambio de contexto y la

aceptación o rechazo por parte del estudiante de las nuevas dinámicas educativas implementadas determinan la decisión de continuar o no haciendo parte de la comunidad educativa, la teoría de "conductas de logro"propuesta por Ethington (1990) plantea que el nivel de aspiraciones y expectativas de éxito del estudiante son elementos fundamentales que explican la deserción, de igual forma Bean and Eaton (2001) determinan que los procesos llevados a feliz término están relacionados con la integración académica y social que tienen los estudiantes, además proponen como estrategia de intervención el acompañamiento académico y psicológico de los estudiantes durante su primer año de estudios superiores.

Por otra parte, un enfoque sociológico hace referencia al estudiante como un ser comunitario en el cual además de sus convicciones personales influyen agentes externos en la toma de sus decisiones así como el entorno y el contexto que lo rodea; Spady (1970) resalta la importancia de los factores externos al estudiante como pueden ser la influencia familiar, las normas establecidas de comportamiento y las directrices establecidas por las IES. Bank et al. (1990) confirman que a mayor interacción de los estudiantes con sus docentes y sus pares académicos es menor el riesgo de abandonar la Universidad, pues al afianzar estos lazos el nivel de compromiso a su sus deberes académicos aumenta considerablemente y se disminuye el riesgo de desertar.

El gobierno nacional durante los últimos años ha encaminado sus acciones hacia los modelos costo - beneficio y a la focalización de subsidios, con los cuales se busca que a través del acompañamiento y de incentivos económicos brindados por el Estado, el estudiante pueda culminar su formación académica y que las instituciones cuenten con recursos financieros destinados al seguimiento de esta problemática. El papel que juegan las IES es resaltado por Kamens (1971) quien afirma que el nivel organizacional mostrado por las instituciones así como su tamaño y la calidad de su personal, influyen en la decisión de los estudiantes de

continuar en el alma máter.

Finalmente el enfoque interaccionista, considera que la deserción estudiantil está ligada a la forma como el estudiante desarrolla sus inteligencias intrapersonales e interpersonales, así como el reconocimiento de su quehacer dentro de una Institución, Tinto (1975) explica el proceso de permanencia del estudiante, de acuerdo a los niveles de ajuste existentes entre el estudiante y el entorno educativo que le rodea, el cual es medible mediante las experiencia académicas y sociales vivenciadas, Bean and Metzner (1985) por su parte hacen una analogía de la vida académica de un estudiante con los modelos de productividad en el ámbito laboral en el cual las variables que describen el ambiente institucional son reemplazadas por variables propias del ambiente laboral los cuales impactan de forma positiva o negativa tanto en las actitudes como en las decisiones tomadas por los individuos en torno a las intenciones de abandonar la academia.

Algunos de los enfoques mencionados anteriormente así como sus máximos representantes a través del tiempo y los factores determinantes de la deserción escolar, son resumidos en la Figura 2.1 elaborada por (Castaño et al., 2006) y empleada en el libro “Deserción estudiantil en la educación superior colombiana”(Ruiz et al., 2009).

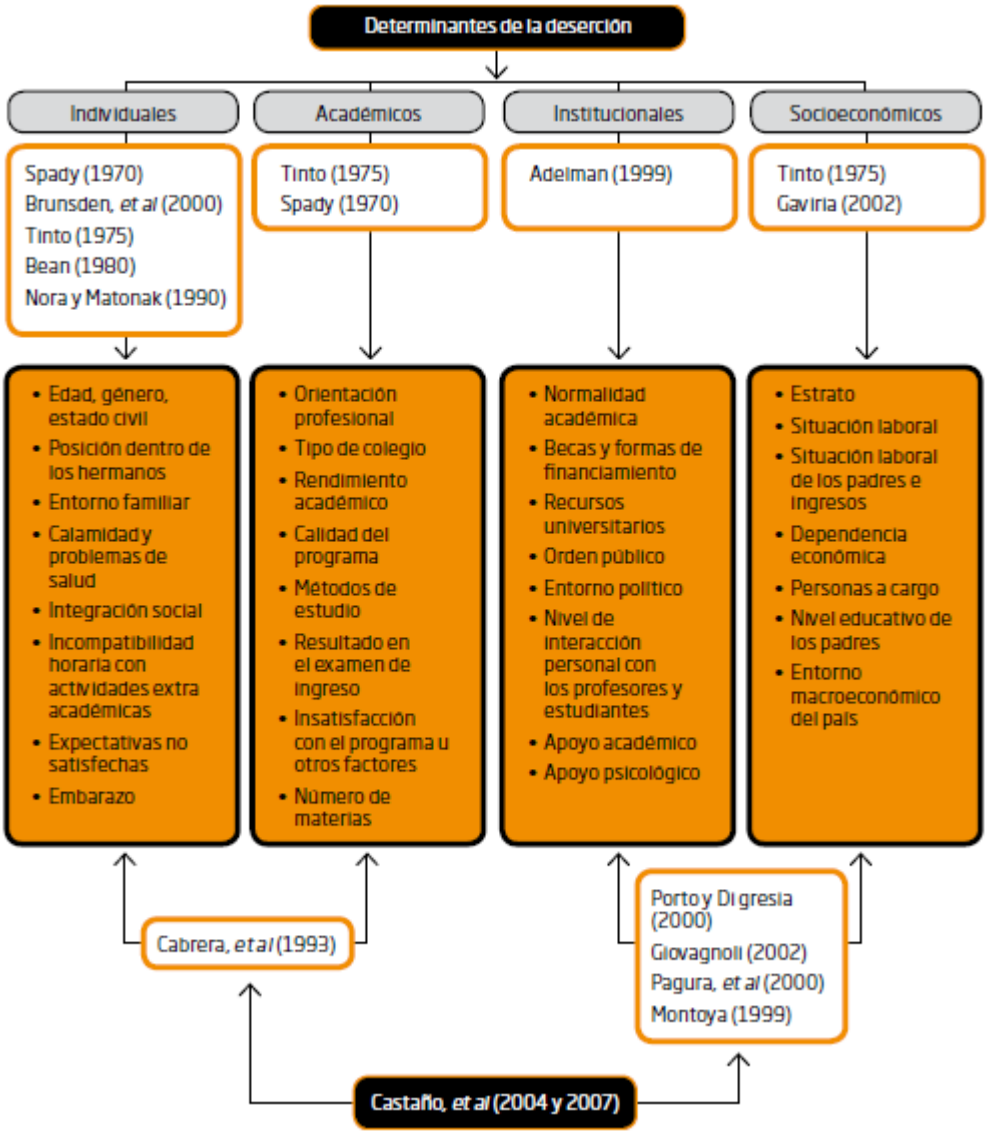

Figura 2.1: Estado del arte de los determinantes de la deserción. (Tomado de Desercion estudiantil en la educación superior colombiana: metodología de seguimiento, diagnóstico y elementos para su prevención)

La deserción universitaria en Colombia, como fue mencionado anteriormente es monitoreada mediante el software SPADIES. A través de este sistema se obtiene información de la deserción por periodo (Figura 2.2) y por cohorte (Figura 2.3) en los diferentes niveles de formación: técnica, tecnológica y universitaria del año 2015. La deserción por periodo es definida como la proporción de estudiantes que no se matriculan en dos semestres consecutivos y deserción por cohorte es la proporción acumulada de estudiantes que ingresaron a primer semestre en un programa académico en un mismo periodo académico (cohorte) y que no registran matrícula en dos o más periodos consecutivos.

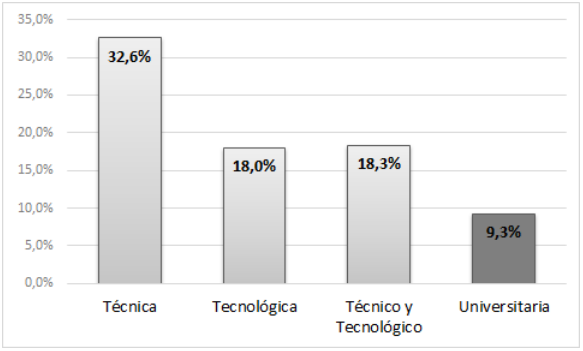

Figura 2.2: Deserción por periodo 2015 (fuente: SPADIES).

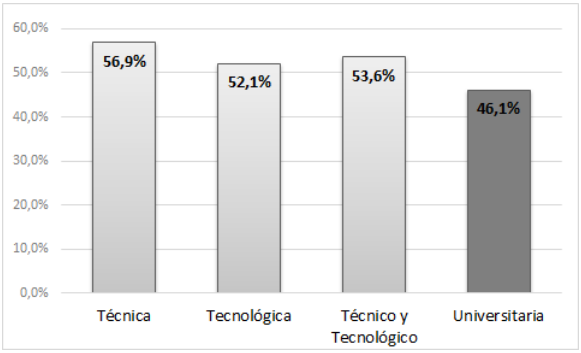

Figura 2.3: Deserción por cohorte 2015 (fuente: SPADIES).

Para el año 2015 el gobierno nacional tenía como meta que a lo más un $9,7\,\%$ de los estudiantes que ingresaran a la educación superior en nivel universitario se convirtieran en desertores, sin embargo, el resultado obtenido es del 9,3 % lo cual determina un aumento significativo en la persistencia académica de los estudiantes que logran ingresar a la universidad, así mismo, los resultados obtenidos para la deserción por cohorte muestra que un 46,1 % de los estudiantes que inician su proceso académico a nivel universitario no logra llevar a feliz término con sus metas propuestas.

El departamento del Quindío presenta una tasa de deserción por cohorte del 49,2 % la cual se encuentra por encima de la obtenida nacionalmente $46,1\,\%$ mientras que la deserción por periodo fue del 7,2 % que está muy por debajo de la media nacional de $9,3\,\%$ como se puede observar en las Figuras 2.4 y 2.5.

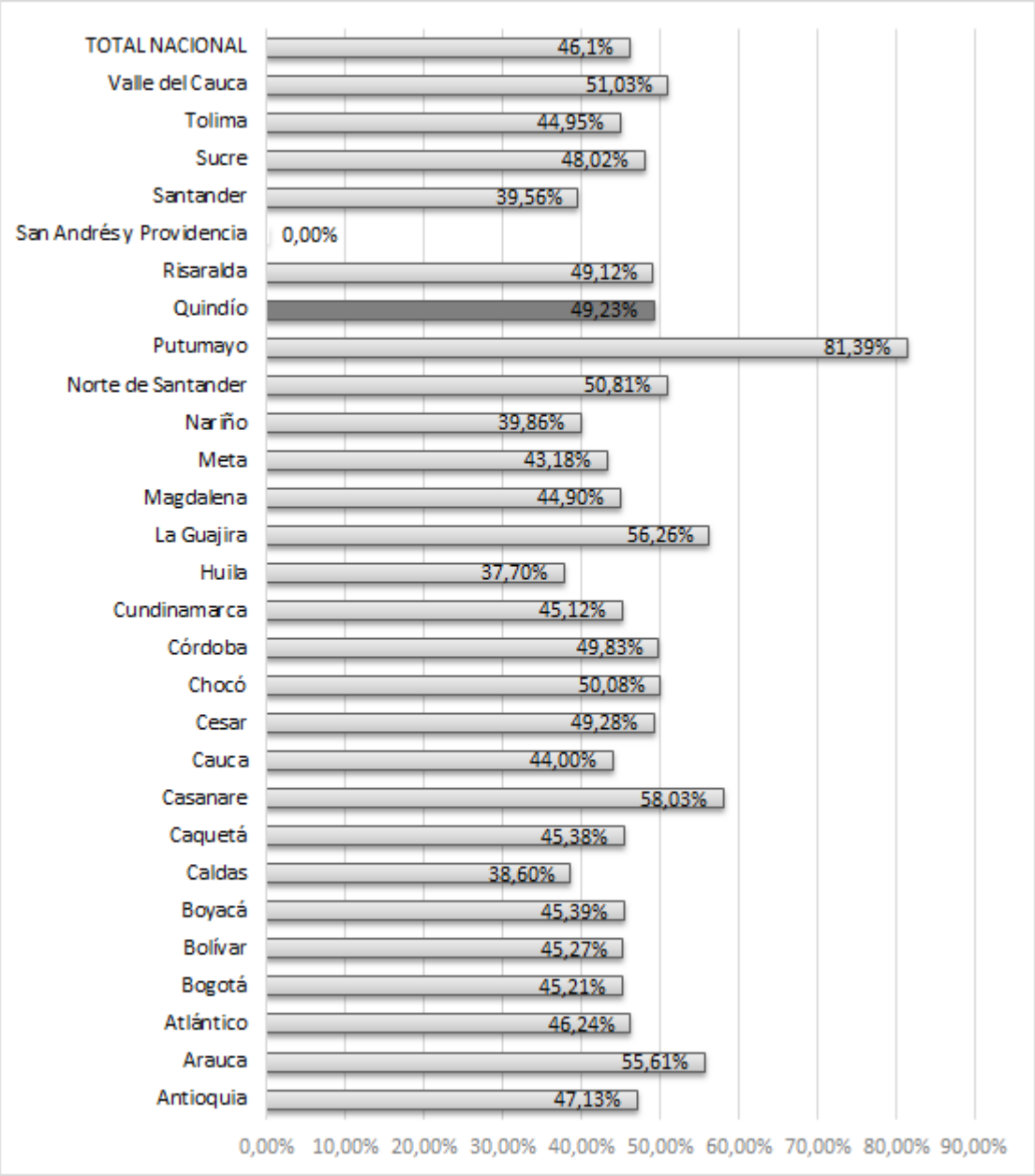

Figura 2.4: Deserción por cohorte y departamentos 2015 (fuente: SPADIES).

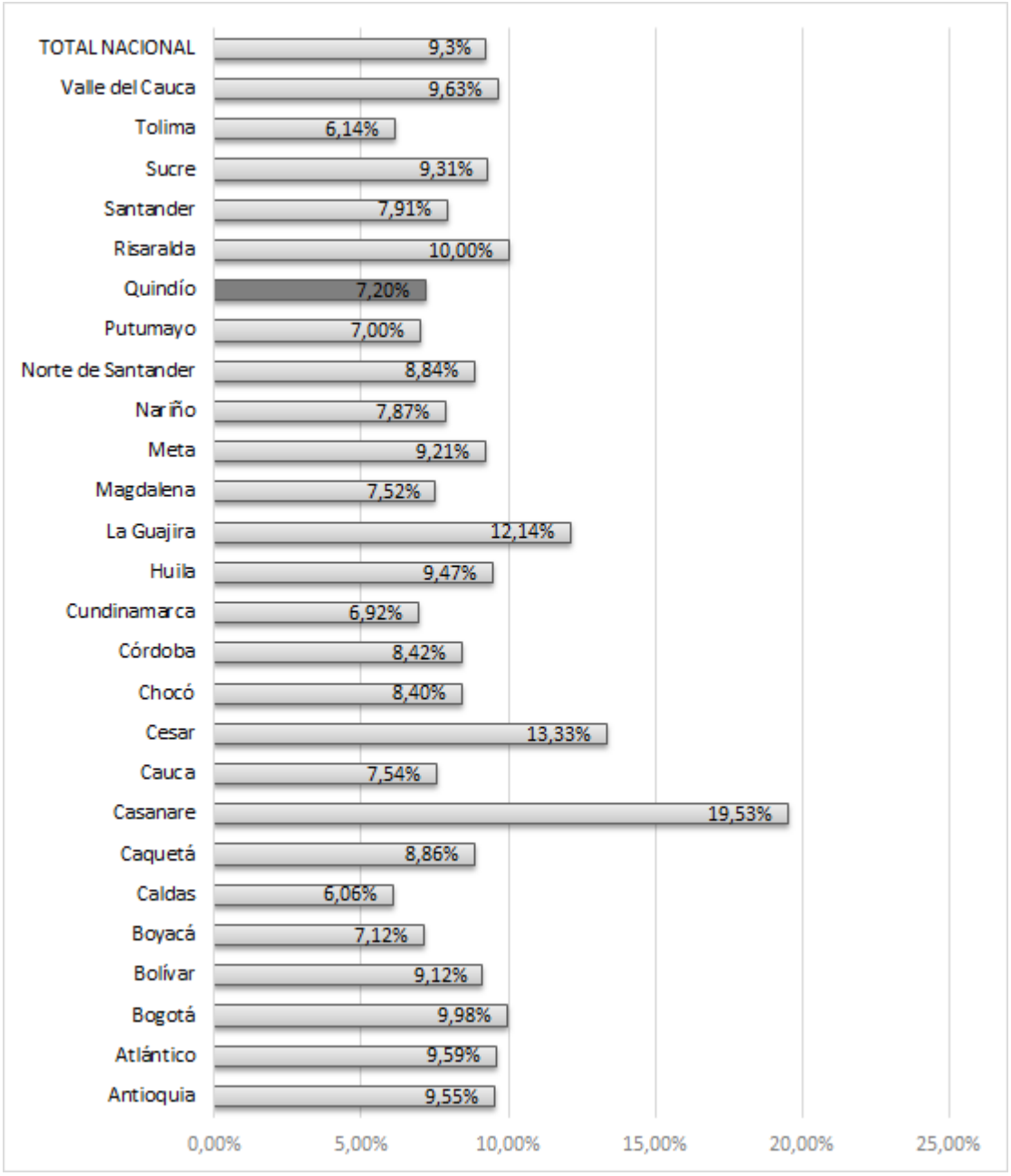

Figura 2.5: Deserción por periodo y departamentos 2015 (fuente: SPADIES).

Capítulo 3

Modelo ZIP

En un modelo ZIP (Lambert, 1992) se tienen dos clases de ceros: los generados por la distribución Poisson que aparecen con probabilidad $1-p$, y un conjunto de ceros extra que aparecen con probabilidad p, se supone que la variable respuesta se distribuye como la mezcla de una distribución de Poisson (λ) y una distribución de masa puntual en cero, con probabilidad de mezcla p. Tanto a p como a λ se les permite depender de covariables a través de funciones de enlace adecuadas.

Para definir el modelo ZIP, considere el vector respuesta $\mathbf{Y} = \{Y_1, Y_2, Y_3, ..., Y_n\}$ de tal forma que sus componentes son independientes y distribuidos de la forma

$$Y_i \sim \begin{cases} 0 & \text{con probabilidad} \quad p_i \\ Poisson \quad (\lambda_i) & \text{con probabilidad} \quad 1-p_i \end{cases}$$

Así, la función de masa de probabilidad de Y_i es

$$P(Y_i = y \mid \lambda_i, p_i) = \begin{cases} p_i + (1-p_i) \quad \exp(-\lambda_i) & \text{si} \quad y = 0 \\ \\ (1-p_i) \quad \frac{\exp(-\lambda_i)\lambda_i{}^y}{y!} & \text{si} \quad y = 1, 2, .. \end{cases} \tag{3.1}$$

En muchas aplicaciones prácticas es común suponer que los parámetros p_i y λ_i dependen de vectores con variables explicativas $\boldsymbol{x_i}$ y $\boldsymbol{z_i}$, respectivamente. En este trabajo, se asumen los enlaces canónicos, es decir

$$\begin{aligned} \log(\lambda_i) &= \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} \quad = \boldsymbol{x_i^T \beta}, \qquad (3.2) \\ \text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) = \gamma_0 + \gamma_1 z_{1i} + ... + \gamma_q z_{qi} \quad = \boldsymbol{z_i^T \gamma}, \qquad (3.3) \end{aligned}$$

en donde tanto $\boldsymbol{\beta}$ como $\boldsymbol{\gamma}$ son parámetros desconocidos de la regresión de dimensión $p \times 1$ y $q \times 1$, respectivamente. La estimación de estos parámetros puede ser realizada usando métodos de máxima verosimilitud como en (Lambert, 1992) (Böhning et al., 1999) (Jones et al., 2001) (Nanjundan and Raveendra Naika, 2012), sin embargo, en este trabajo será utilizada la estimación Bayesiana a través de Cadenas de Markov vía Montecarlo (MCMC) implementado mediante software como OpenBUGS® o WinBUGS (Ntzoufras, 2011).

Función de verosimilitud

Consideremos ahora una muestra de las observaciones $(y_1, x_1, z_1), ..., (y_n, x_n, z_n)$ de n observaciones independientes, donde cada respuesta observada se denota por y_i. En este caso, la función de verosimilitud para el vector de los parámetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})^T$ dada la muestra tiene la forma

$$\begin{aligned} L(\theta) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) \quad &= \prod_{y_i=0} [p_i + (1-p_i) \quad \exp(-\lambda_i)] \times \prod_{y_i>0} \left[(1-p_i)\frac{\exp(-\lambda_i) \quad \lambda^{y_i}}{y_i!}\right] \\ &= L_1(\theta) \times L_2(\theta), \end{aligned}$$

donde

$$
\begin{aligned}
L_1(\theta) &= \prod_{i:y_i=0} \left(\quad p_i + (1-p_i)\exp(-\lambda_i) \quad \right) && (3.4)\\
&= \prod_{i:y_i=0} \left(\frac{\exp(z_i^T\gamma)}{1+\exp(z_i^T\gamma)} + \frac{\exp(-\exp(x_i^T\beta))}{1+\exp(z_i^T\gamma)} \right) \quad \text{y} && (3.5)\\
L_2(\theta) &= \prod_{i:y_i>0} (1-p_i) \frac{\exp(-\lambda_i) \quad \lambda^{y_i}}{y_i!} && (3.6)\\
&= \prod_{i:y_i>0} \left(\frac{1}{1+\exp(z_i^T\boldsymbol{\gamma})} \frac{\exp(-\exp(x_i^T\beta))\exp(x_i^T\beta y_i)}{y_i!} \right). && (3.7)
\end{aligned}
$$

Las Ecuaciones (3.5) y (3.7) resultan de despejar λ_i y p_i de las Ecuaciones (3.2) y (3.3), respectivamente y reemplazarlas en las Ecuaciones (3.4) y (3.6)

Así, la función de log-verosimilitud $l(\theta)$ puede ser expresada como:

$$
l(\boldsymbol{\theta}) = \log\left(L(\boldsymbol{\theta})\right) = \log(L_1(\boldsymbol{\theta})) + \log(L_2(\boldsymbol{\theta})) = l_1(\boldsymbol{\theta}) + l_2(\boldsymbol{\theta}) \tag{3.8}
$$

donde

$$
\begin{aligned}
l_1(\boldsymbol{\theta}) &= \sum_{i:y_i=0} \log\left(\frac{\exp(z_i^T\gamma) + \exp(-\exp(x_i^T\beta))}{1+\exp(z_i^T\gamma)} \right)\\
&= \sum_{i:y_i=0} \log\Big(\exp(z_i^T\gamma) + \exp(-\exp(x_i^T\beta))\Big) - \sum_{i:y_i=0} \log\Big(1+\exp(z_i^T\gamma)\Big)
\end{aligned}
$$

y

$$
\begin{aligned}
l_2(\boldsymbol{\theta}) &= \sum_{i:y_i>0} \log\left(\frac{1}{1+\exp(z_i^T\boldsymbol{\gamma})} \frac{\exp\Big(-\exp(x_i^T\boldsymbol{\beta})\Big)\exp(x_i^T\boldsymbol{\beta}y_i)}{y_i!} \right)\\
&= \sum_{i:y_i>0} \Big(-\exp(x_i^T\boldsymbol{\beta}) + x_i^T\boldsymbol{\beta}y_i\Big) - \sum_{i:y_i>0} \log\Big(1+\exp(z_i^T\boldsymbol{\gamma})\Big) - \sum_{i:y_i>0} \log\left(y_i!\right)
\end{aligned}
$$

3.1. Estimación de los Parámetros

Un paso fundamental en el desarrollo de la metodología Bayesiana es la asignación de las distribuciones a priori para los parámetros del modelo. En este trabajo se usarán distribuciones normales independientes, para los parámetros de la regresión $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$, en que $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ y $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_q)$, específicamente se tiene que $\beta_i \sim N(0, 100)$ y $\gamma_j \sim N(0, 100)$, $i = 1, \ldots, p$; $j = 1, \ldots, q$. Por tanto, la distribución conjunta a priori puede ser escrita como

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\,\pi(\boldsymbol{\gamma}), \tag{3.9}$$

en que

$$\pi(\boldsymbol{\beta}) = \prod_{i=1}^{p}\left(\frac{1}{\sqrt{2\pi}\sqrt{100}}\quad \exp\left(\frac{-\beta_i^2}{2\times 100}\right)\right) = \left(\frac{1}{200\pi}\right)^{\frac{p}{2}}\exp\left(\frac{-\sum_{i=1}^{p}\beta_i^2}{200}\right) \quad \text{y}$$

$$\pi(\boldsymbol{\gamma}) = \prod_{i=1}^{q}\left(\frac{1}{\sqrt{2\pi}\sqrt{100}}\quad \exp\left(\frac{-\gamma_j^2}{2\times 100}\right)\right) = \left(\frac{1}{200\pi}\right)^{\frac{q}{2}}\exp\left(\frac{-\sum_{i=1}^{q}\gamma_i^2}{200}\right)$$

Luego la ecuación (3.9) puede escribirse como

$$\pi(\boldsymbol{\theta}) = \left(\frac{1}{200\pi}\right)^{\frac{p+q}{2}}\quad \exp\left(-\frac{1}{200}\left(\sum_{i=1}^{p}\beta_i^2 + \sum_{j=1}^{q}\gamma_j^2\right)\right)$$

Luego, la distribución a posteriori es obtenida a través de

$$\pi(\theta|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) = l(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}) \tag{3.10}$$

la cual puede ser escrita para $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$, respectivamente como

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) = l_1(\theta)\,l_2(\theta)\,\pi(\beta),$$
$$\pi(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) = l_1(\theta)\,l_2(\theta)\,\pi(\gamma)$$

Teniendo en cuento lo expresado en las ecuaciones (3.8) y (3.9) se determina que la ecuación (3.10) puede escribirse como:

$$\pi\left(\theta|\boldsymbol{y},\boldsymbol{x},\boldsymbol{z}\right) = \left[\sum_{i:y_i=0}\log\Big(\exp(z_i^T\gamma)+\exp(-\exp(x_i^T\beta))\Big) - \sum_{i:y_i=0}\log\Big(1+\exp(z_i^T\gamma)\Big)\right.$$
$$\left. + \Big(-\exp(x_i^T\boldsymbol{\beta}) + x_i^T\boldsymbol{\beta}y_i\Big) - \sum_{i:y_i>0}\log\Big(1+\exp(z_i^T\boldsymbol{\gamma})\Big) - \sum_{i:y_i>0}\log\left(y_i!\right)\right]$$
$$\times\left[\left(\frac{1}{200\pi}\right)^{\frac{p+q}{2}}\exp\left(-\frac{1}{200}\left(\sum_{i=1}^{p}\beta_i^2+\sum_{j=1}^{q}\gamma_j^2\right)\right)\right]$$

Debido a la forma que presenta la distribución a posteriori de los parámetros de interés, en este trabajo se usa un enfoque Bayesiano para estimar estos parámetros. Los pasos de los algoritmos MCMC fueron implementados usando la libreria RBugs (Ligges et al., 2010), la cual conecta a R con el software OpenBUGS®. Para cada parámetro se generaron dos cadenas independientes cada una de tamaño 10000, en donde las primeras 5000 se consideraron burn-in y de las siguientes 5000 (con un salto 10) se obtuvieron las estimaciones de cada uno de ellos. La convergencia fue monitoreada a través del factor de reducción de escala Brooks-Gelman-Rubin (Brooks et al., 2004), $\hat{R}$ disponible en la libreria R coda (Cowles and Carlin, 1996).

Capítulo 4

Descripción de los datos analizados

Los datos que motivaron este trabajo provienen de un estudio de cohorte realizado en la Universidad del Quindío, específicamente se obtiene información de 288 estudiantes que ingresaron a primer semestre durante el año 2012 y que abandonaron sus estudios luego de permanecer entre uno y cuatro semestres. Estos estudiantes no presentaron matricula en dos semestres consecutivos con lo que, según el criterio del MEN y de la Universidad del Quindío, se consideran desertores.

La información recolectada contiene los resultados obtenidos en las pruebas Saber 11 y en las pruebas BADyG (Batería de Aptitudes Diferenciales y Generales) (Díez and Marín, 2009) (Yuste et al., 2001), estas últimas son utilizadas para evaluar el desarrollo de algunos procesos cognitivos. Además se obtuvo información relacionada con aspectos académicos , familiares y económicos.

En la Tabla (4.1) se presentan detalladamente las variables utilizadas en el ajuste del modelo ZIP. A continuación se hace una descripción de las componentes evaluadas en las pruebas BADyG así como un análisis estadístico de los resultados obtenidos de la muestra obtenida.

RESULTADOS PRUEBA SABER 11	
GENERAL	Promedio ponderado
ÁREAS EVALUADAS	Biología
	Sociales
	Filosofía
	Física
	Inglés
	Lenguaje
	Matemáticas
	Química
PRUEBA BADyG	
INTELIGENCIA GENERAL	Completar oraciones
	Resolución de problemas
	Encajar figuras
RAZONAMIENTO LÓGICO	Relaciones analógicas
	Series numéricas
	Matrices lógicas
MEMORIA	Memoria auditiva
	Memoria Visual
ATENCIÓN	Atención
RAPIDEZ Y EFICACIA	Rapidez
	Eficacia
ASPECTOS ACADÉMICOS	Jornada
	Nombre del programa
	Tipo de colegio
	Cantidad de materias registradas
	Cantidad de materias aprobadas
ASPECTOS SOCIO-FAMILIARES	Edad
	Genero
	Procedencia geográfica
ASPECTOS SOCIO-ECONÓMICOS	Estrato
	Valor de la matricula

Tabla 4.1: Factores que pueden llegar a influir en la deserción universitaria

4.1. Prueba BADy-G

Durante las dos primeras semanas de clase, se les aplica a los estudiantes que ingresan a primer semestre en la modalidad presencial, la prueba BADyG con la que se evalúan aspectos como la inteligencia general, la capacidad de razonamiento lógico, la memoria, la atención y finalmente la rapidez y la eficacia. Estos aspectos se evalúan a través de algunos componentes detallados a continuación.

Inteligencia general (IG): Esta inteligencia está determinada por las capacidades de razonamiento general del estudiante, así como el dominio de habilidades cognitivas como:

- Completar oraciones (Sv): con la cual se observa el dominio del vocabulario y los conocimientos previos del estudiante.
- Resolución de problemas(Sn): mide la rapidez en la ejecución de operaciones matemáticas al momento de resolver problemas numéricos determinados.
- Encajar figuras (Se): determina las habilidades espaciales de acuerdo a la percepción intuitiva o racional del entorno físico y de los objetos que hay en él; mediante la realización giros y encajes espaciales.

La capacidad de razonamiento lógico (Rl): el razonamiento es un proceso mediante el cual partiendo de uno o más juicios de valor se determina la posibilidad o falsedad en la ocurrencia en situaciones específicas y se evalúan mediante:

- Relaciones analógicas(Rv): esta prueba mide el razonamiento a través de las habilidades de comprensión verbal, mediante la lectura critica de diferentes textos.
- Series numéricas (Rn): en esta prueba se evalúa el razonamiento numérico, la cual mide la capacidad de cálculo mental en el manejo de operaciones básicas matemáticas así como el dominio de los diferentes conjuntos numéricos.

- Matrices Lógicas (Re): mide el razonamiento espacial, mediante la resolución de secuencias en las cuales intervienen figuras geométricas.

Ademas existen sub-pruebas adicionales que miden aspectos diferenciadores de la personalidad tales como:
Memoria: ésta se mide a través de dos pruebas:

- Memoria auditiva (Ma): evalúa la capacidad para retener información a través de un relato oral.
- Memoria visual (Mv): evalúa la distinción visual ortográfica de palabras, que dependerá de la retentiva a largo plazo.

Atención: la capacidad de atención se evalúa mediante la rapidez en la discriminación visual de gráficos y dibujos.
Rapidez y eficacia Son dos variables añadidas en esta prueba en las cuales, la rapidez mide el tiempo empleado en contestar las seis pruebas,mientras que la eficacia es una relación entre el numero de aciertos con relación a la totalidad de las respuestas.

El promedio de los resultados obtenidos por los estudiantes desertores en estas pruebas en general y clasificados de acuerdo al tipo de colegio de donde provienen son resumidas en la Tabla 4.2. A partir de esta, se puede concluir que, a excepción de los resultados obtenidos por los estudiantes provenientes de colegios privados cuando se evalúa la Inteligencia General, estos son superados o igualados por los resultados obtenidos por los estudiantes que provienen de colegios oficiales en el restante de componentes evaluados.

TIPO DE COLEGIO	IG	Rl	Rv	Rn	Re	Sv	Sn	Se	Ma	Mv	Ra
PRIVADO	91,00	59,17	42,04	36,75	32,75	39,04	17,92	44,71	32,50	38,71	126,83
OFICIAL	81,17	59,09	43,04	40,94	36,45	44,20	21,22	51,12	32,92	44,57	101,38
TOTAL	81,99	59,09	42,95	40,59	36,14	43,77	20,95	50,59	32,88	44,08	103,50

Tabla 4.2: Resultados pruebas BADyG teniendo en cuenta el tipo de colegio de procedencia.

Cuando se comparan los resultados de las pruebas BADyG en cada uno de los componentes entre mujeres y hombres se observan claras diferencias siendo que, con excepción de los resultados obtenidos cuando se evalúa la inteligencia general, los hombres superan a las mujeres como puede ser observado en la Figura 4.1.

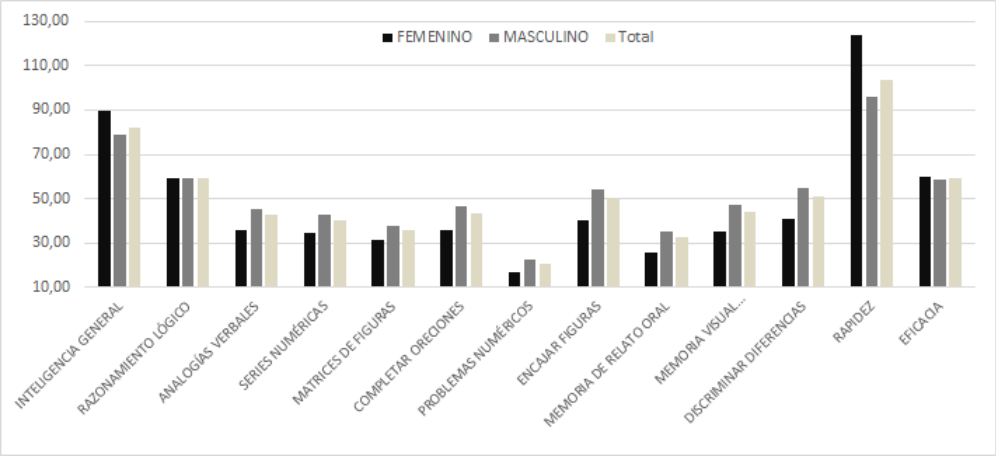

Figura 4.1: Resutados obtenidos en la prueba BADyG teniendo en cuenta el género.

A partir de la muestra analizada se tiene también que el promedio obtenido en las pruebas Saber 11 es de $51,5$ de 100 posibles con una desviación estándar de $6,97$. No se observan claras diferencias entre el promedio obtenido por mujeres y hombres $51,41$ y $51,54$, respectivamente, aunque la variación en las mujeres es menor como se puede observar en la Tabla 4.3.

	PROM	BIOLOGÍA	C.SOCIALES	FILOSOFÍA	FÍSICA	INGLÉS	LENGUAJE	MATEMÁTICAS	QUÍMICA
Media F.	51,41	51,62	49,62	48,19	48,66	45,87	53,57	53,03	49,18
SD	6,62	8,43	7,51	10,66	8,90	10,11	7,86	12,16	7,17
Media M.	51,54	50,91	48,56	45,70	50,99	46,02	51,29	55,09	47,76
SD	7,11	8,99	9,40	9,96	9,14	8,02	8,95	10,09	7,74
Media	51,50	51,10	48,85	46,37	50,36	45,98	51,91	54,53	48,14
SD	6,97	8,83	8,93	10,20	9,12	8,62	8,71	10,70	7,60

Tabla 4.3: Resultados en las pruebas Saber 11 general y discriminado por sexo y área. SD es la desviación estándar.

En la Figura 4.2 se comparan los resultados obtenidos por los estudiantes desertores en las pruebas Saber 11 provenientes de instituciones oficiales y privadas. En esta figura se puede observar que a excepción del comportamiento obtenido en física, los estudiantes que provenían de colegios privados obtuvieron un mayor puntaje que aquellos que lo hacían de colegios oficiales. En la Figura 4.3 se resumen otras variables analizadas; se puede decir por ejemplo que el $91,66\,\%$ de los desertores de la Universidad provienen de instituciones educativas oficiales mientras que sólo el $8,34\,\%$ provienen de instituciones privadas, lo cual evidencia las diferencias en la preparación y adaptación a entornos académicos de mayor rigor.

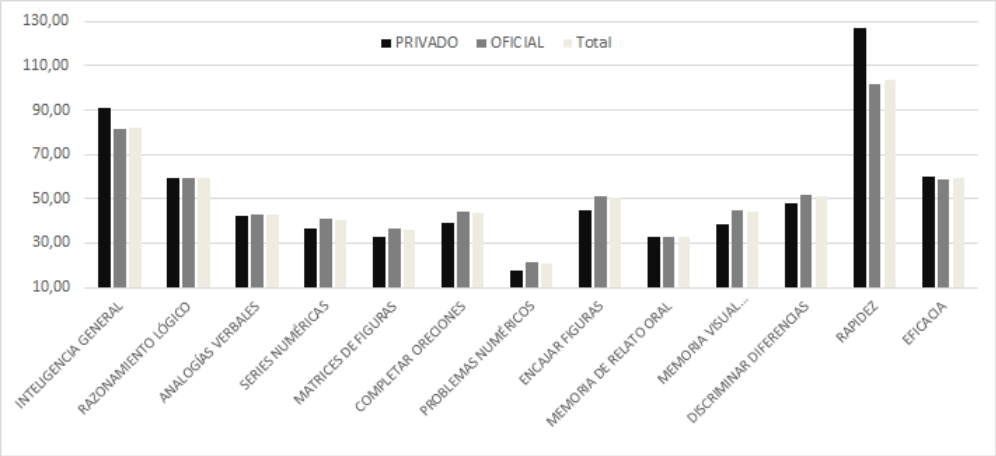

Figura 4.2: Comparación del comportamiento de algunas variables por tipo de institución.

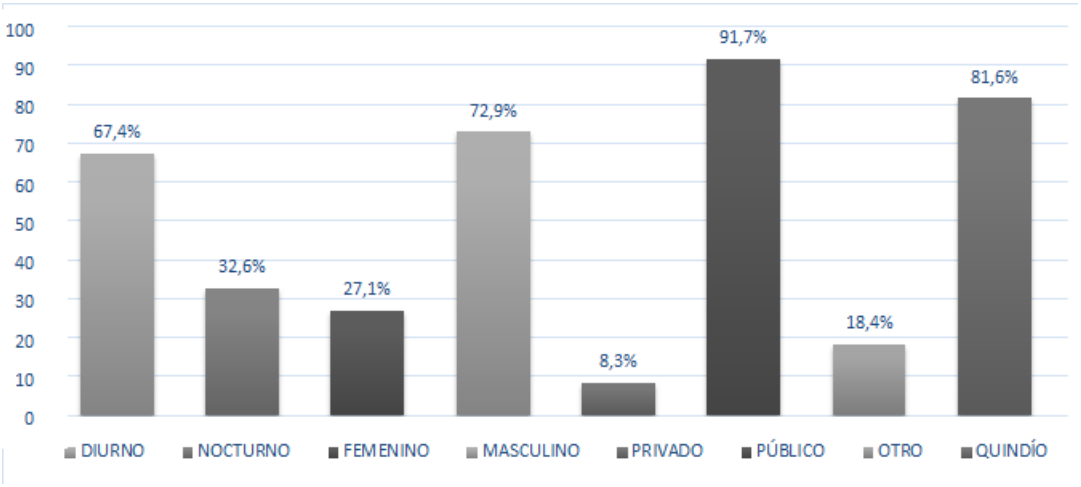

Figura 4.3: Distribución de los estudiantes desertores según la jornada, el genero, el tipo de colegio y la procedencia geográfica.

Otras variables analizadas permiten concluir que el $72,9\,\%$ de los estudiantes desertores son hombres entre 19 y 38 años con una edad promedio que no supera los 23 años, en su mayoría pertenecientes al departamento del Quindío y a programas diurnos.

El $58,7\,\%$ de los estudiantes desertores son provenientes de la facultad de Ingeniería, seguidos de la facultad de Ciencias Básicas y Tecnológicas $24,3\,\%$ y luego se encuentran los provenientes de facultad de Educación $12,8\,\%$. También se tiene que el valor pagado por concepto de matricula de 264 de los 288 estudiantes considerados como desertores tempranos no supera el salario mínimo legal vigente; teniendo en cuenta que el colegio del cual provienen determina el valor a pagar; el $91,66\,\%$ pertenecen al sector oficial.

En términos del número de materias registradas y aprobadas por los estudiantes desertores se puede decir que en promedio registraron 9 materias durante su permanencia en el programa académico, de las cuales en promedio aprobaron 4. También se puede decir que el número máximo de materias registradas por un

estudiante durante su estadía en la Universidad fue de 22, de las cuales aprobó 14 de ellas, mientras que el mínimo de materias registradas fue de 2 retirándose con sólo una de ellas aprobada.

4.2. Variables explicativas de la deserción empleadas en el modelo

La deserción universitaria tiene múltiples causas enmarcadas en factores individuales, económicos , académicos e institucionales y por eso debe ser analizada desde diferentes enfoques que posibiliten la detección de los elementos determinantes así como su posible tratamiento, sin embargo, las variables a tener en cuenta dentro de cada grupo de factores, dependen de la información disponible en las fuentes de información.

Para la implementación del modelo de regresión ZIP expresado en la Ecuación (3.1) se tienen en cuenta los resultados obtenidos por los estudiantes en las pruebas Saber 11 en general y en las áreas de ciencias sociales, inglés, lenguaje y matemáticas, también se tienen los resultados con respecto a la inteligencia general la cual sintetiza las seis pruebas básicas evaluadas a través de las pruebas BADyG, así como el valor obtenido en la memoria del relato oral y la memoria visual ortográfica, los cuales brindan información del desarrollo cognitivo e intelectual actual de cada estudiante.

Otros factores que se utilizaron son la edad y el género. También se incluyó el valor de la matrícula, el cual puede ser pensado como determinante de la decisión de dar por terminado el proceso académico de un estudiante sin la culminación del programa académico al cual pertenece. Finalmente pero no menos importantes

están los factores académicos propios de la Institución Educativa reflejados en la Jornada (diurna o nocturna), el número de materias registradas por semestre y el tipo de colegio en el cual el estudiante terminó sus estudios de educación media que puede ser diurna o nocturna.

Capítulo 5

Aplicación del modelo ZIP

En la aplicación del modelo ZIP se considera como variable respuesta el número de materias aprobadas. Debido a la naturaleza de esta variable (conteo) podría ser ajustado un modelo de regresión de Poisson, sin embargo, como puede ser observado en la Figura 5.1, se presenta un exceso de ceros, lo que hace no recomendable el uso de este modelo y justifica el uso del modelo de regresión ZIP.

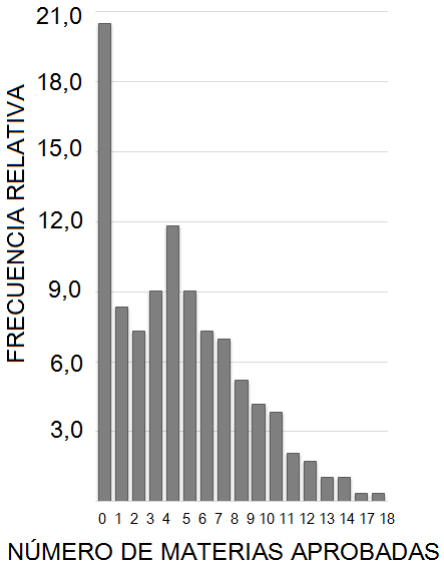

Figura 5.1: Cantidad de materias aprobadas antes de tomar la desición de abandonar los estudios universitarios.

En la Tabla 5.1 se puede observar que de 217 estudiantes (obtenido como la suma de los totales de las primeras dos columnas) que desertaron en primer y segundo semestre, 164 (obtenido como la suma de los totales de las primeras cinco filas) lo hicieron teniendo aprobadas, como máximo, 4 materias. Luego es posible concluir que el número de materias aprobadas efectivamente puede ser visto como un indicador de deserción temprana.

	Semestre				
Y	1°	2°	3°	4°	Total
0	58	1	0	0	59
1	22	2	0	0	24
2	15	6	0	0	21
3	16	7	3	0	26
4	20	11	3	0	34
5	14	9	3	0	26
6	4	7	10	0	21
7	0	11	8	1	20
8	0	5	7	3	15
9	0	7	1	4	12
≥ 10	0	2	16	12	30
Total	149	68	51	20	288

Tabla 5.1: Número de materias aprobadas en los cuatro primeros semestres por los estudiantes considerados como desertores.

Con la información recolectada con respecto a los los aspectos socio-económicos, socio-familiares, académicos y los resultados obtenidos en las pruebas saber 11 y las BADyG, se procede a ajustar el modelo de regresión ZIP presentado en la Ecuación (3.1) en donde se asume como variable respuesta Y_i al número de materias aprobadas por el i-ésimo estudiante en el momento de abandonar el programa académico; y donde los parámetros p_i y λ_i son modelados a través de

$$\begin{aligned} \log\lambda_i &= \boldsymbol{x_i^T\beta} \quad , \\ \text{logit}p_i &= \boldsymbol{z_i^T\gamma}, \end{aligned}$$

en que $\boldsymbol{x_i^T} = \boldsymbol{z_i^T}$=(1, memoria auditiva$_i$, memoria visual$_i$, inteligencia general$_i$, total de materias registradas$_i$, jornada$_i$, genero$_i$, edad$_i$, valor de la matricula$_i$, promedio ponderado saber 11$_i$, puntaje en sociales$_i$, puntaje en ingles$_i$, puntaje en lenguaje$_i$, puntaje en matemáticas$_i$, tipo de colegio$_i$). El uno es utilizado para acomodar el intercepto y $\boldsymbol{\beta^T} = (\beta_1, \beta_2, \cdots, \beta_{15})$ y $\boldsymbol{\gamma^T} = (\gamma_1, \gamma_2, \cdots, \gamma_{15})$.

Las estimaciones de los parámetros del modelo ZIP así como sus correspondientes desviaciones estándar son presentados en la Tabla 5.2.

Parámetro	**Estimación**	**SD**	**Parámetro**	**Estimación**	**SD**
Intercepto	0,450	0,337	Intercepto	-3,645	5,061
Memoria auditiva	-0,027	0,046	Memoria auditiva	-0,253	0,336
Memoria visual	0,021	0,046	Memoria visual	0,626	0,345
Inteligencia general	0,050	0,042	Inteligencia general	-0,149	0,301
Materias registradas	0,464	0,029	Materias registradas	-2,151	0,469
Jornada*	1,377	0,298	Jornada*	-1.678	3,506
Genero*	-0,394	0,402	Genero	2,141	2,106
Edad	0,057	0,03	Edad	0,026	0,243
Valor de la matrícula	0,019	0,034	Valor de la matrícula	-0,755	0,664
Promedio Saber 11	0,045	0,059	Promedio Saber 11	-0,072	0,301
Prueba de Sociales	0,052	0,04	Prueba de Sociales	0,347	0,305
Prueba de Inglés	0,014	0,034	Prueba de Inglés	-0,421	0,316
Prueba de Lenguaje	0,038	0,037	Prueba de Lenguaje	-0,153	0,279
Prueba de Matemáticas	-0,062	0,048	Prueba de Matemáticas	-0,371	0,363
Tipo de colegio	0,122	0,151	Tipo de colegio	-2,040	10,65

Tabla 5.2: Estimaciones y desviaciones estándar (SD) de los parámetros en el modelo de regresión ZIP que modelan la media (columna izquierda) y los que modelan el exceso de ceros (columna derecha). Las variables significativas aparecen con *.

Una vez estimado el vector de parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ es posible calcular la probabilidad estimada de que un estudiante abandone su programa académico con cero, uno, dos, ... y hasta diez materias aprobadas usando el modelo (3.1) de la siguiente forma

$$\begin{aligned}
\widehat{P(Y_i = 0)} &= \quad \widehat{p}_i + (1 - \widehat{p}_i) \quad \exp(-\widehat{\lambda}_i) \\
\widehat{P(Y_i = 1)} &= (1 - \widehat{p}_i) \, \frac{\widehat{\lambda}_i \quad \exp(-\widehat{\lambda}_i)}{1!} \\
\widehat{P(Y_i = 2)} &= (1 - \widehat{p}_i) \, \frac{\widehat{\lambda}_i^2 \quad \exp(-\widehat{\lambda}_i)}{2!} \\
&\vdots \\
\widehat{P(Y_i = 10)} &= (1 - \widehat{p}_i) \, \frac{\widehat{\lambda}_i^{10} \quad \exp(-\widehat{\lambda}_i)}{10!},
\end{aligned}$$

Donde $\widehat{\lambda}_i = \exp \quad (\widehat{\beta}_1 + \quad \widehat{\beta}_2 \text{ memoria auditiva}_i + \cdots + \quad \widehat{\beta}_{15} \text{ tipo de colegio}_i)$ y $\widehat{p}_i = \dfrac{\exp(a_i)}{1 + \exp(a_i)}$ con $a_i = \quad \widehat{\gamma}_1 + \quad \widehat{\gamma}_2 \text{ memoria auditiva}_i + \cdots + \quad \widehat{\gamma}_{15} \text{ tipo de colegio}_i$, en que los valores $\widehat{\beta}_k$ y $\widehat{\gamma}_k$ con $k = 1, 2, ..., 15$ son presentados en la Tabla 5.2.

La máxima probabilidad estimada indicará cuál es el número de materias aprobadas con el que desertará el estudiante, por lo tanto, si éste tiene una alta probabilidad de desertar al trascurrir 4 o menos materias aprobadas, podría ser identificado como un desertor temprano.

Con el propósito de evaluar el modelo se hace el cálculo de su poder predictivo, esto es la capacidad que tiene para predecir qué estudiantes abandonarán tempranamente sus estudios basados en el número de materias aprobadas, el cual se compara con los valores observados. Estos resultados son presentados en la Tabla 5.3

	Semestre				
Y	1°	2°	3°	4°	Total
0	96	2	0	0	98
2	18	0	0	0	18
3	33	15	0	0	48
4	2	19	4	0	25
5	0	18	7	0	25
6	0	8	14	1	23
7	0	4	6	1	11
8	0	1	10	2	13
9	0	1	6	3	10
10	0	0	4	13	17
Total	149	68	51	20	288

Tabla 5.3: Número de materias aprobadas antes de retirarse, estimadas a través del modelo de regresión ZIP.

En esta tabla se observa que de los 98 estudiantes identificados a través del modelo como aquellos que abandonarían sus estudios con 0 materias aprobadas, 96 desertaron efectivamente en primer semestre mientras que los 2 restantes lo hicieron en el segundo semestre; así mismo, se puede observar que de los 18 estudiantes que fueron identificados por el modelo como aquellos que abandonarían sus estudios con 2 materias aprobadas, todos desertaron en primer semestre. Además, también es importante notar en esta tabla que el modelo estima con más de 10 materias aprobadas aquellos estudiantes que abandonan sus estudios universitarios en tercer semestre o más.

En forma general, en este trabajo se considera que un estudiante desertará tempranamente cuando, a través del modelo ZIP, se estime que aprobará 4 o menos materias antes de retirarse de la universidad y usando este criterio es posible detectar el 100 % de los estudiantes que abandonarán sus estudios en primer semestre y una gran parte de los que lo harán en segundo semestre. Así que es posible concluir que el modelo ZIP puede ser utilizado como una herramienta es-

tadística que permite identificar estudiantes que tienen alta chance de abandonar tempranamente sus estudios universitarios y una vez identificados, las instituciones de educación superior deberían implementar estrategias que conlleven a que estos estudiantes obtengan un título profesional y así lograr impactar no sólo su calidad de vida sino también la de su entorno familiar.

Capítulo 6

Estudio de Simulación

Para analizar el efecto que sobre los coeficientes de la regresión tiene el exceso de ceros en un modelo de regresión de Poisson se realizaron dos estudios de simulación en los cuales se evaluaron el Error Cuadrático Medio (ECM), el sesgo relativo absoluto (SR) y el porcentaje de cobertura (CP) del verdadero valor del parámetro.

En el primer caso se generaron 100 muestras para diferentes valores de n de una distribución de Poisson con una media λ_i modelada a través de $\lambda_i = \exp \beta_1 + \beta_2 X_i$ donde $\beta_1 = \beta_2 = 2$ y la variable X_i fue generada a partir de una distribución Uniforme(0,1) y n tomó los valores 50, 100, 150 y 200 indicando diferentes tamaños de muestras.

El siguiente paso consistió en ubicar los ceros utilizando una distribución Uniforme (0,1) y una distribución de Bernoulli con parámetro p_i de la siguiente manera. Se calcula p_i a través de $p_i = \dfrac{\exp(\gamma_1 + \gamma_2 Z_i)}{1 + \exp(\gamma_1 + \gamma_2 Z_i)}$, donde $\gamma_1 = -1$, $\gamma_2 = 1$ y Z_i se generó a partir de una distribución Normal(0,2). Si el valor de la uniforme generada es menor o igual que p_i entonces se ubica un cero en el i-ésimo valor generado inicialmente con distribución de Poisson, en caso contrario el valor ge-

nerado no es cambiado.

Como fue mencionado anteriormente se generaron 100 conjuntos de datos para cada valor de n y fueron ajustados los modelos ZIP y Poisson. En este caso $\theta = \{\beta_1, \beta_2, \gamma_1, \gamma_2\}$ y θ_s un elemento de θ y se calculó $\text{ECM}(\theta_s) = \frac{1}{100}\sum_{i=1}^{100}(\hat{\theta}_{is} - \theta_s)^2$, $\text{SR}=\frac{1}{100}\sum_{i=1}^{100}\left|\frac{\hat{\theta}_{is}}{\theta_s} - 1\right|$ y finalmente el Porcentaje de cobertura calculado como el porcentaje de veces en el que el verdadero valor del parámetro pertenecía al intervalo del 95 % de credibilidad.

Los resultados obtenidos a través de este estudio de simulación son presentados en la Figura 6.1 y la Tabla 6.1.

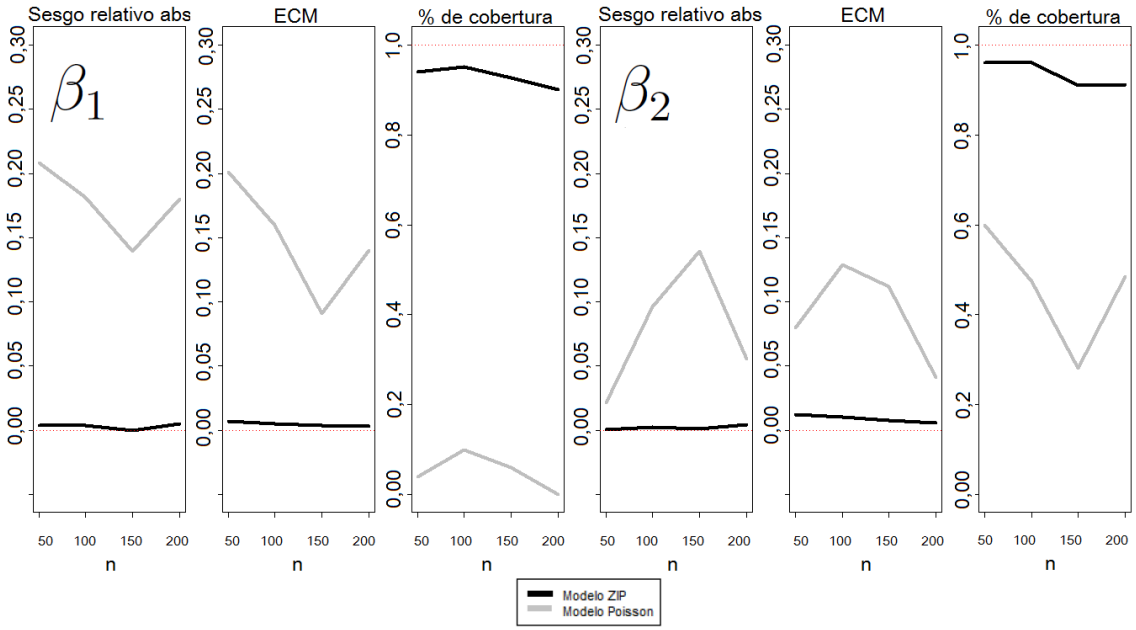

Figura 6.1: Sesgo relativo absoluto, Error cuadrático medio y porcentaje de cobertura para β_1 y β_2 después de ajustar los modelos ZIP (línea negra) y Poisson (línea gris)

		Sesgo relativo absoluto				Error cuadrático medio				Porcentaje de cobertura			
		n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
ZIP	β_1	0,004	0,004	0,000	0,005	0,007	0,005	0,004	0,003	0,940	0,950	0,927	0,901
	β_2	0,000	0,002	0,001	0,004	0,012	0,010	0,008	0,006	0,960	0,960	0,909	0,911
	γ_1	0,205	0,091	0,002	0,007	0,297	0,104	0,058	0,043	0,940	0,990	0,964	0,950
	γ_1	0,251	0,122	0,052	0,020	0,222	0,073	0,050	0,022	0,940	0,921	0,918	0,960
Poisson	β_1	0,208	0,181	0,140	0,180	0,201	0,160	0,090	0,140	0,040	0,099	0,060	0,000
	β_2	0,021	0,096	0,140	0,055	0,080	0,129	0,112	0,041	0,600	0,475	0,280	0,485

Tabla 6.1: SR, ECM y CP para β_1, β_2 , γ_1 y γ_2 después de ajustar los modelos ZIP y Poisson.

Los valores obtenidos para β_1 y β_2 con respecto al sesgo en el modelo ZIP están muy cercanos a cero para todas muestras tomadas, la diferencia es significativa con respecto a los obtenidos para los mismos parámetros por el modelo de Poisson.

Las predicciones realizadas por el modelo ZIP cuentan con una mayor precisión, ya que el ECM de β_1 y β_2 son considerablemente cercanos a cero para todas las muestras tomadas que van desde n=50 hasta n=200, en comparación con los valores obtenidos por el modelo ZIP. Esto evidencia la importancia de tener en cuenta la cantidad de observaciones con valor cero presentes en los datos al momento de realizar una inferencia sobre estos.

En el segundo estudio de simulación , el parámetro p_i se considera constante para todo i. Específicamente se consideraron los valores $p = 0,2$ y $p = 0,3$ siguiendo el esquema anteriormente descrito. Posteriormente, fueron ajustados los modelos ZIP y Poisson. Los resultados obtenidos con $p = 0{,}2$ se observan en la Figura 6.2 (parte superior) y en la parte inferior se presentan los resultados obtenidos con $p = 0,3$. Los resultados obtenidos también se presentan en las Tablas

6.2 y 6.3.

		Sesgo relativo absoluto				Error cuadrático medio				Porcentaje de cobertura			
		n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
ZIP	β_1	0,001	0,006	0,002	0,002	0,007	0,003	0,002	0,002	0,960	0,960	0,990	0,933
	β_2	0,004	0,009	0,000	0,003	0,017	0,005	0,004	0,004	0,92	0,99	0,99	0,94
	p	0,049	0,031	0,049	0,006	0,003	0,002	0,001	0,001	0,940	0,930	0,920	0,973
Poisson	β_1	0,124	0,104	0,110	0,110	0,087	0,057	0,087	0,055	0,310	0,220	0,090	0,047
	β_2	0,014	0,018	0,006	0,003	0,091	0,037	0,091	0,021	0,600	0,650	0,660	0,593

Tabla 6.2: Sesgo relativo absoluto, ECM y CP para β_1 y β_2 después de ajustar los modelos ZIP y Poisson con p =0,2.

		Sesgo relativo absoluto				Error cuadrático medio				Porcentaje de cobertura			
		n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
ZIP	β_1	0,003	0,001	0,004	0,003	0,007	0,003	0,003	0,003	0,970	0,960	0,950	0,940
	β_2	0,006	0,001	0,005	0,003	0,015	0,008	0,007	0,005	0,97	0,95	0,95	0,96
	p	0,019	0,003	0,019	0,006	0,004	0,002	0,001	0,001	0,930	0,970	0,960	0,920
Poisson	β_1	0,165	0,176	0,182	0,172	0,168	0,146	0,168	0,130	0,320	0,070	0,010	0,000
	β_2	0,033	0,010	0,001	0,021	0,172	0,076	0,172	0,040	0,500	0,500	0,630	0,550

Tabla 6.3: Sesgo relativo absoluto, ECM y CP para β_1 y β_2 después de ajustar los modelos ZIP y Poisson con $p = 0{,}3$

El porcentaje de cobertura de los parámetros β_1 y β_2 en las simulaciones realizadas para el modelo de Poisson no superan el 70 %, mientras que los resultados para el modelo ZIP nunca están por debajo del del 90 %.

Como se puede observar en los estudios de simulación realizados en ese trabajo y mediante los resultados presentados en las Figuras 6.1, 6.2 y en las Tablas 6.1 y 6.3 se puede concluir que el modelo ZIP supera al modelo de Poison cuando se compara el sesgo relativo absoluto, el ECM y finalmente el porcentaje de cobertura cuando la presencia de observaciones con valor cero excede la esperada por el modelo de Poisson.

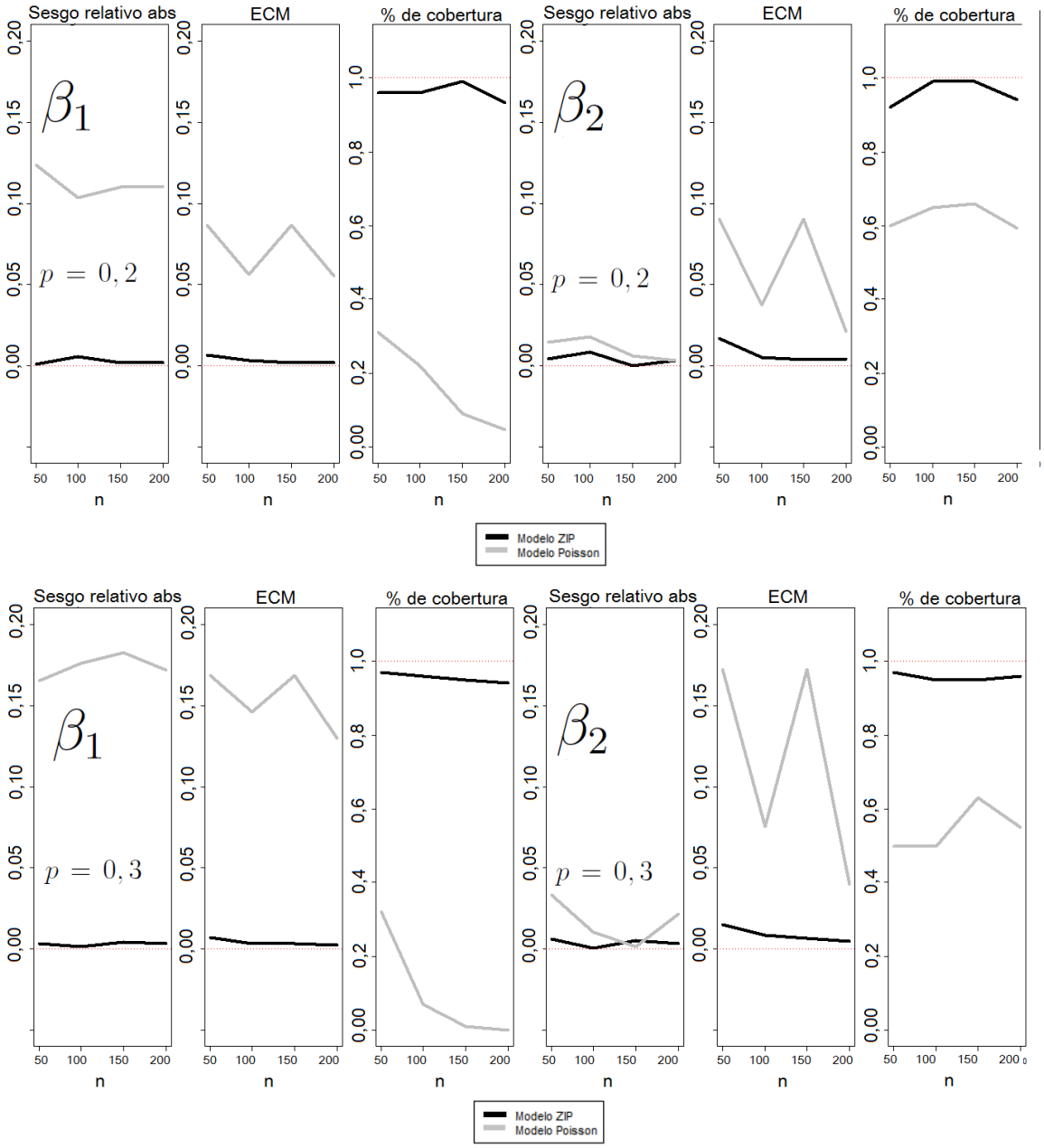

Figura 6.2: Sesgo relativo absoluto, Error cuadrático medio y porcentaje de cobertura para β_1 y β_2 después de ajustar los modelos ZIP (línea negra) y Poisson (línea gris) con $p = 0{,}2$ (figura superior) y con $p = 0{,}3$ (figura inferior).

Capítulo 7

Conclusiones y Recomendaciones

- La deserción universitaria constituye un gran problema para el sistema educativo y para la comunidad en general, puesto que incide de manera directa en el desarrollo social, económico y político de una nación que desea aminorar las brechas de desigualdad social.

- La deserción universitaria es un problema multicausal que puede ser visto como una respuesta a diferentes factores personales, sociales, económicos o académicos (Galvis et al., 2014) y su análisis puede ser direccionado desde diferentes enfoques para definir la variable respuesta lo que daría origen a una diversidad de análisis estadísticos; en este trabajo se utiliza el número de materias aprobadas por un estudiante como un indicador de deserción temprana y su análisis se hace a través del modelo ZIP.

- El modelo ZIP, como fue mostrado en este trabajo, puede utilizarse para identificar posibles desertores tempranos y esta información puede servir de apoyo a la administración de la Universidad para dirigir adecuadamente las estrategias de intervención, tanto académicas como económicas de las que se dispone y así evitar que muchos estudiantes abandonen sus estudios de educación superior y logren su objetivo de obtener un título profesional.

- La función de riesgo construida con base en el modelo ZIP puede ser utilizada como una herramienta estadística que permite identificar a los estudiantes que corren mayor riesgo de abandonar tempranamente sus estudios universitarios y una vez identificados, las IES podrían implementar estrategias asertivas que contribuyan a aumentar la tasa de graduación.

- En la literatura han sido utilizados diferentes modelos para el análisis de la deserción tales como modelos de regresión mixtos (Little, 1995) el modelo probit (Montmarquette et al., 2001), modelos de análisis de sobrevivencia (Ligges et al., 2010), el modelo logístico multinomial (Stratton et al., 2008), ecuaciones estructurales (Khan et al., 2013) redes neuronales (Siri, 2015), técnicas de minería de datos (Torres et al., 2016), entre otros; sin embargo, no se tiene conocimiento del uso de modelos de regresión ZIP en el estudio de la deserción.

Bibliografía

Ajzen, I. and Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological bulletin*, 84(5):888.

Attinasi, L. (1986). Getting in: Mexican american students perceptions of their college-going behavior with implications for their freshman year persistence in the university. In *ASHE 1986 Annual Meeting Paper, San Antonio, TX.(ERIC No. ED 268 869)*.

Bank, B. J., Slavings, R. L., and Biddle, B. J. (1990). Effects of peer, faculty, and parental influences on students' persistence. *Sociology of education*, pages 208–225.

Bean, J. and Eaton, S. B. (2001). The psychology underlying successful retention practices. *Journal of College Student Retention: Research, Theory & Practice*, 3(1):73–89.

Bean, J. P. and Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of educational Research*, 55(4):485–540.

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(2):195–209.

Brooks, S., King, R., and Morgan, B. (2004). A bayesian approach to combining animal abundance and demographic data. *Animal Biodiversity and Conservation*, 27(1):515–529.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Castaño, E., Gallón, S., Gómez, K., and Vásquez, J. (2006). Análisis de los factores asociados a la deserción y graduación estudiantil universitaria. *Lecturas de economía*, (65):9–36.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Díez, E. M. and Marín, M. Á. C. (2009). Descripción de los factores medidos por la batería badyg-m y su estudio como variables de intervención educativa. *Revista Iberoamericana de Educación*, 6(49).

Ethington, C. A. (1990). A psychological model of student persistence. *Research in higher Education*, 31(3):279–293.

Galvis, D., Hurtado, L., Garcia, M., and Mendez, R. (2010). *La desercion estudiantil en la Universidad del Quindío, diagnóstico y estrategia de intervención*. Editorial Universidad del Quindío.

Galvis, D. M., Bandyopadhyay, D., and Lachos, V. H. (2014). Augmented mixed beta regression models for periodontal proportion data. *Statistics in medicine*, 33(21):3759–3771.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Jones, B. L., Nagin, D. S., and Roeder, K. (2001). A sas procedure based on mixture models for estimating developmental trajectories. *Sociological methods & research*, 29(3):374–393.

Kamens, D. H. (1971). The college "charter" and college size: Effects on occupational choice and college attrition. *Sociology of education*, pages 270–296.

Khan, Y. A., Ahamad, Z., and Kousar, S. (2013). Factors influencing academic failure of university students. *International Journal of Educational Administration and Policy Studies*, 5(5):79–84.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K., and Sturtz, S. (2010). Brugs 0.5: Openbugs and its r/s-plus interface brugs. r package.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Montmarquette, C., Mahseredjian, S., and Houle, R. (2001). The determinants of university dropouts: a bivariate probability model with sample selection. *Economics of education review*, 20(5):475–484.

Müller, P. (1991). *A generic approach to posterior integration and Gibbs sampling*. Purdue University, Department of Statistics.

Nanjundan, G. and Raveendra Naika, T. (2012). Asymptotic comparison of method of moments estimators and maximum likelihood estimators of parameters in zero-inflated poisson model. *Applied mathematics*, 3(6):610–616.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, volume 698. John Wiley & Sons.

Ruiz, C. G., Muriel, D. M. D., Gallego, J. F., Velez, E. C., Gomez, S. G., and Portilla, K. G. (2009). *Desercion estudiantil en la educacion superior colombiana: metodologia de seguimiento, diagnóstico y elementos para su prevención*. Ministerio de Educacion Nacional.

Siri, A. (2015). Predicting students dropout at university using artificial neural networks. *Italian Journal of Sociology of Education*, 7(2).

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1):64–85.

Stratton, L. S., O'Toole, D. M., and Wetzel, J. N. (2008). A multinomial logit model of college stopout and dropout behavior. *Economics of education review*, 27(3):319–331.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1):89–125.

Torres, C. Z., Ramos, C. A., and Moraga, J. L. (2016). Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos. *Ciencia Amazónica (Iquitos)*, 6(1):73–84.

Yuste, C., Martínez, R., and Galve, J. (2001). Batería de aptitudes diferenciales y generales (badyg, battery of diferencial aptitudes of intelligence, iq test).

Apéndice A

Código BUGS para implementar el modelo ZIP

Código BUGS para implementar el modelo ZIP usando covariables para modelar p

```
model
{
Cte<-10
for(j in 1:n)
{
 zeros[j]       <- 0
 zeros[j]       ~  dpois(zeros.means[j])
 zeros.means[j] <- -lPoisInf[j]+Cte
 lPoisInf[j]     <- log(p[j]*equals(Y[j],0)+(1-p[j])*fdPois[j])
 fdPois[j]      <- exp(-lambda[j]+Y[j]*log(lambda[j])-loggam(Y[j]+1))
 log(lambda[j]) <- beta[1]+beta[2]*x[j,1]
 logit(p1[j])   <- gamma[1]+gamma[2]*x[j,2]
 p[j]           <- max(0.00001,min(0.9999,p1[j]))
}
```

```
Priors of parameters  beta and gamma
for (i in 1:2)
  {
   beta[i]   ~ dnorm(0,0.001)
   alpha[i]  ~ dnorm(0,0.001)
  }
}
```

Código BUGS para implementar el modelo ZIP usando p constante

```
model
{
Cte<-10
for(j in 1:n)
{
 zeros[j]       <- 0
 zeros[j]       ~  dpois(zeros.means[j])
 zeros.means[j] <- -lPoisInf[j]+Cte
 lPoisInf[j]    <- log(p*equals(Y[j],0)+(1-p)
 fdPois[j]      <- exp(-lambda[j]+Y[j]*log(lambda[j])-loggam(Y[j]+1))
 log(lambda[j]) <- beta[1]+beta[2]*x[j,1]
}
Priors of parameters beta and  p
p ~ dunif(0,1)
for (i in 1:2)
  {
   beta[i]   ~ dnorm(0,0.001)
  }
}
```

Apéndice B

Inferencia Bayesiana

A diferencia de la inferencia frecuentista en la cual los valores de probabilidad se entienden como la frecuencia relativa de ocurrencia de un suceso a largo plazo, la inferencia Bayesiana supone que las cantidades desconocidas (parámetros) son variables aleatorias en vez de constantes, y que los datos, una vez observados, son fijos en vez de aleatorios. Por esta razón la estimación realizada mediante un enfoque Bayesiano no consiste en encontrar estimadores puntuales de los parámetros de interés, si no en encontrar una distribución de probabilidad completa para dichos parámetros.

El proceso de estimación de un vector de parámetros o parámetro $\theta \in \Theta$, desde una perspectiva bayesiana se basa en la distribución *a posteriori* de θ denotada por $\pi(\theta|y)$, donde $\boldsymbol{Y^T} = (Y_1, \ldots, Y_n)$ está conformado por n variables aleatorias independientes cada una de ellas con función de densidad de probabilidad (fdp) $f(y|\theta)$. Una distribución *a posteriori* $\pi(\theta|y)$ es obtenida a partir del Teorema de Bayes combinando una función de verosimilitud $L(\theta|y)$ dada por $\prod_{i=1}^{n} f(Y_i|\theta)$ y una distribución *a priori* $\pi(\theta)$ asignada a los parámetros. Esto permite resumir el conocimiento previo y la evidencia recolectada en una distribución de probabilidad

denotada por

$$\pi(\theta|y) = \frac{L(\theta|y)\pi(\theta)}{f(y)}, \tag{B.1}$$

en que $f(y)$ es la fdp marginal de y, que se obtiene como $\int_{\Theta} L(\theta|y)\pi(\theta)d\theta$ y que se conoce como constante de normalización.

Un estimador de θ es obtenido por medio de una característica de la distribución *a poteriori* como la media o la mediana. Estas características son llamadas, respectivamente, como media o mediana *a poteriori*.

Una vez obtenida la distribución *a poteriori* de θ se constituye la base de la inferencia Bayesiana y es posible utilizar métodos computacionales como métodos Monte Carlo basados en cadenas de Markov (MCMC), generan muestras de esta distribución sin necesidad de simular directamente de dicha distribución.

Estos métodos se basan en la construcción de una cadena de Markov ergódica, aperiodica e irreducible cuya distribución estacionaria es la distribución de interés. Por lo tanto, después de un número de iteraciones, los valores generados en un resultado de muestra en la distribución deseada.

Los métodos MCMC más comunes son el muestreador de Gibbs y el algoritmo Metropolis-Hastings, los cuales se explican brevemente a continuación.

Algoritmo Metropolis-Hastings

Este algoritmo fue propuesto inicialmente por Nicholas Metropolis y extendido por W. Keith Hasting. Metropilis propuso y empleó este algoritmo por primera vez para el caso específico de la distribución de Boltzmann (Metropolis et al., 1953). Posteriormente Hasting presentó una versión para casos más generales (Hastings, 1970), este algoritmo se ha utilizado ampliamente en física, sin embargo, era po-

co conocido por los estadísticos hasta hace poco. Los trabajos de (Müller, 1991) y (Tierney, 1994) fueron fundamentales para exponer el valor de este algoritmo y estimular el interés de los estadísticos en su uso. A través de este algoritmo, es posible generar muestras de una distribución de probabilidad necesitando únicamente f.d.p. de interés $\pi(\theta|y)$, lo que representa una ventaja en el caso de la estadística bayesiana ya que a menudo calcular la constante de normalización puede ser en una tarea difícil.

Este algoritmo utiliza una distribución a partir de la cual es fácil generar muestras llamada *Kernel de transición*, denotada por $r(\cdot|\cdot)$. Kernel de transición es $\pi(\theta|y)$ Es la distribución estacionaria de la cadena de Markov generada. En la descripción del algoritmo, supongamos θ Es un escalar. El algoritmo esta determinado por los siguiente pasos:

1. Determinar un valor inicial para θ, denotado por $\theta^{(0)}$.

2. En la k-ésima iteración, se genera un nuevo valor θ^* a partir de $r(\theta^{(k-1)}|\cdot, y)$

3. Este valor se acepta con probabilidad $\alpha(\theta^{(k-1)}, \theta^*)$ dada por

$$\min\left\{1, \frac{\pi(\theta^*|y)r(\theta^*|y, \theta^{(k-1)})}{\pi(\theta^{k-1}|y)r(\theta^{(k-1)}|y, \theta^*)}\right\}.$$

   Generando una variable aleatoria (v.a.) U con una distribución uniforme $(0, 1)$, el nuevo valor es aceptado si $U < \alpha$, esto es $\theta^{(j)} = \theta^*$ de lo contrario $\theta^{(j)} = \theta^{(j-1)}$.

4. Repetir los pasos 2 e 3 para $k = 1, 2, \ldots$

Como en el muestreador de Gibbs, el proceso comienza con un valor arbitrario, y luego de ser iterado un número de veces y después de un *burn-in*, e genera la distribución de las observaciones sobre la distribución de interés. En este caso, el kernel de transición r genera un valor (candidato) que será aceptada en función del valor de α.

Muestreador de Gibbs

El muestreador de Gibbs propuesto por Geman and Geman (1984) es uno de los métodos MCMC más conocidos; por medio de éste se generan muestras aleatorias de la distribución de interés de forma indirecta mediante el siguiente algoritmo.

Suponiendo la distribución condicional $\theta_i|\mathbf{y}, \theta_{(-i)}$, $i = 1, \ldots, l$, el algoritmo puede ser implementado mediante los siguientes pasos:

1. Determinar valores iniciales para los componentes de $\boldsymbol{\theta}$, denotado por $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_l^{(0)})$.

2. Generar la k-ésima iteración como sigue

   * $\theta_1^{(k)} \sim \boldsymbol{\theta}_1^{(k)}|\theta_{(-1)}^{(k-1)}$.
   * $\theta_2^{(k)} \sim \boldsymbol{\theta}_2^{(k)}|y, \theta_1^{(k)}, \theta_3^{(k-1)}, \ldots, \theta_l^{(k-1)}$.
   * $\theta_3^{(k)} \sim \boldsymbol{\theta}_3^{(k)}|y, \theta_1^{(k)}, \theta_2^{(k)}, \theta_4^{(k-1)}, \ldots, \theta_l^{(k-1)}$.
   * $\vdots$
   * $\theta_l^{(k)} \sim \boldsymbol{\theta}_l^{(1)}|y, \theta_{(-l)}^{(k)}$.

3. Repetir para $k = 1, 2, \ldots$.

El proceso comienza con un valor arbitrario, el cual se itera un número de veces y después de un *burn-in*, se generan observaciones de la distribución de interés, una de las ventajas más importantes de este algoritmo es que todas las simulaciones son aceptadas, y en cada transición se obtiene un punto diferente de la cadena. Otros detalles del muestreador de Gibbs se pueden encontrar en (Gelfand and Smith, 1990) y (Casella and George, 1992), entre otros.