PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

APLICACIÓN DE UN MODELO DE RIESGOS COMPETITIVOS BAYESIANO

Tesis para obtener el grado académico de Maestro en Estadística que presenta:

Erick Dennis Saavedra Palacios

Asesor:

Victor Giancarlo Sal y Rosas Celi

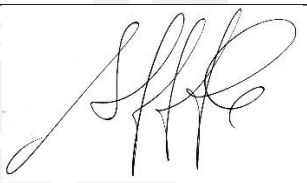
Lima, 2023

Informe de Similitud

Yo, Víctor Giancarlo Sal y Rosas Celi, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis/el trabajo de investigación titulado “Aplicación de un modelo de riesgos competitivos bayesiano”, del autor Erick Dennis Saavedra Palacios, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 3%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 15/11/2023.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 01 de febrero de 2024.

Apellidos y nombres del asesor / de la asesora: Sal y Rosas Celi, Victor Giancarlo	
DNI: 40361284	Firma : 
ORCID: 0000-0001-8636-7142	

Dedicatoria

Para mi padre Hermes Abel Saavedra Hipolito por haber puesto todo su esfuerzo y dedicación en darnos una educación y por ser el maestro de toda la vida.

Lo logramos Papá.

Y para mi familia por aguantarme en todos esos momentos en que me encerraba para poder avanzar con mi proyecto de tesis.

Agradecimentos

En primer lugar a dios por darme el conocimiento necesario para poder resolver las cuestiones resueltas en esta tesis, a mi asesor Giancarlo Sal y Rosas por darme la oportunidad de resolver estas cuestiones y brindarme su apoyo en el transcurso de su elaboración, a mis profesores de la maestría por todo el conocimiento brindado en cada clase que me permitieron desarrollar esta tesis y a mi familia y amigos por aguantarme en todo este proceso y por sus animos para culminarla.

Resumen

En el presente trabajo se presenta y discute el modelo de riesgos competitivos bayesiano propuesto por Vallejos y Steel (2017). Dentro del análisis se incluyó un estudio de simulación en donde se comparó los resultados de aplicar el modelo frecuentista con respecto al bayesiano, confirmando la eficiencia de este último con respecto al anterior. Finalmente, se aplicó este modelo a la base de datos de alumnos ingresantes a la Pontificia Universidad Católica del Perú entre los años 2004 a 2012. El resultado de la aplicación mostró como única variable significativa a si el alumno ingreso por la primera opción con respecto al haber ingresado por la vía tradicional la cual aumentaba la probabilidad de graduación y reducía la probabilidad de abandono.

Palabras-clave: Riesgos Competitivos, Deserción estudiantil.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones Preliminares

En muchos aspectos académicos que involucran diferentes disciplinas, el interés por conocer el momento de ocurrencia de algún evento en particular puede ser de mucha importancia. Desde estimar el tiempo de vida de un paciente que presenta una enfermedad terminal o estimar el tiempo que transcurre desde que un estudiante ingresa a la universidad hasta su eventual deserción del centro académico (Trujillo y Raúl, 2016).

La literatura relacionada a la estimación de estos tiempos de falla es conocida bajo el nombre de análisis de supervivencia y ha sido muy extensa en las diferentes disciplinas (Hosmer et al., 2011). Los modelos convencionales asumen un evento de interés (ej. la muerte de un paciente); sin embargo, existen diferentes situaciones en que el análisis demanda explorar más de un evento en cuestión. Este tipo de modelos suelen referirse como modelos de supervivencia en presencia de riesgos competitivos (Hosmer et al., 2011).

Nuestro particular interés está enfocado en los modelos que se han venido desarrollando en materia de la deserción en estudiantes universitarios. Algunos estudios han tratado de modelar la deserción estudiantil con los métodos estándar de supervivencia, es decir, planteando como único evento la deserción estudiantil; sin embargo, bajo este método, los estudiantes que lograron graduarse son tratados como datos censurados (Murtaugh et al., 1999).

Los modelos de riesgos competitivos son más apropiados cuando pueden ocurrir tipos de eventos múltiples y cada evento esta relacionado a diferentes mecanismos. Es más natural pensar que los datos estudiantiles se pueden modelar de esta forma dado que el estudiante universitario esta condicionado al evento de abandonar la universidad de manera voluntaria, de manera involuntaria y de graduarse.

Por otro lado, la mayor parte de la literatura relacionada a riesgos competitivos se ha centrado en tiempos de supervivencia continuos (Crowder, 1996); sin embargo, en el contexto de los resultados universitarios es más natural pensar el tiempo como una medida discreta, dado que los registros se presentan en los semestres académicos (Vallejos y Steel, 2017; Neethling, 2015).

Una particularidad de los datos relacionados a los estudios de deserción estudiantil es el comportamiento de las tasas de riesgo las cuales no tienen una tendencia definida, es decir, el evento que un estudiante se gradúe no se observa en todos los tiempos por la estructura misma del sistema educativo, lo que implica que el riesgo sea cero en los primeros semestres. Este problema puede generar que los rangos de los tiempos de supervivencia no tengan un

gran solapamiento (separación cuasi-completa), generando estimaciones sesgadas, por lo que se propone realizar la estimación bajo el método bayesiano (Vallejos y Steel, 2017).

1.2. Objetivos

El objetivo general de la tesis es presentar y discutir el modelo de riesgos competitivos bajo la perspectiva bayesiana utilizada por Vallejos y Steel (2017) y aplicarlo para los datos de alumnos ingresantes a la Pontificia Universidad Católica del Perú (PUCP) entre los años 2004 y 2012, esto implica reconocer los comportamientos de las funciones de riesgos de los eventos propuestos e identificar las prioris necesarias para su estimación.

Se incluye un análisis de robustez de calibración propuesta, incorporando otras priori para los parámetros de los log-odds basales; asmismo, se analizan las propiedades del modelo y se realiza una implementación computacional del modelo. De manera específica:

- Presentar la formulación del modelo de riesgos competitivos bayesiano introducida por Vallejos y Steel (2017), mostrando los supuestos detrás de las distribuciones a prioris seleccionadas para la implementación del modelo a los datos de la Pontificia Universidad Católica de Chile.
- Presentar la particularidad de los datos recogidos por la PUCP y mostrar si aún son validos los supuestos utilizados por Vallejos y Steel (2017).
- Implementar el modelo mencionado a la base de datos de la PUCP considerando las distribuciones a prioris necesarias para su modelado a través de un software libre .
- Interpretar los resultados del modelo y brindar información acerca de los mecanismo detrás de la deserción estudiantil voluntaria e involuntaria y graduación de los estudiantes de la universidad, incluyendo tanto el evento como el tiempo del suceso.

1.3. Organización del Trabajo

En el capítulo 2 se presenta y discute el modelo frecuentista de riesgos competitivos con tiempo discreto (Cox, 1972). En el capítulo 3 se define y se presenta el modelo de riesgos competitivos bayesiano presentado y discutido por Vallejos y Steel (2017). En el capítulo 4 se desarrolla un estudio de simulación a fin de mostrar la recuperación de parámetros para el modelo propuesto. En el capítulo 5 se analizan los resultados de aplicar el modelo presentado en el ámbito de la deserción estudiantil en la PUCP. Finalmente se presentan conclusiones y sugerencia al presente trabajo en el capítulo 6.

Capítulo 2

Modelo de odds proporcionales para riesgos competitivos

Los modelos de riesgos competitivos han sido utilizados en muchas disciplinas, en donde el objetivo de la investigación no esta relacionado a la ocurrencia de un único evento, sino a multiples eventos que compite entre sí. Es decir, el investigador esta interesado en la distribución del tiempo de falla de uno de estos eventos en la precencia de los demás. Asimsimo, dado el diseño de la investigación se define un tiempo máximo de seguimiento de los individios para el desarrollo de los eventos, por el cual los individuos que cruzan este umbral de tiempo máximo o se retiran de la investigación por alguna razón diferente al evento seran tomados como datos censurados, es decir, inviduos a los cuales se le desconoce si tuvieron el evento o no.

Un ejemplo común de riesgos competitivos es donde un paciente puede morir de diferentes causas tal como cancer, enfermedad al corazón, complicaciones de diabetes (Takam et al., 2009; Wolbers et al., 2009; Clarke et al., 2004). Asimismo, en la literatura correspondiente al análisis de la deserción estudiantil también se ha usado mucho los modelos de riesgos competitivos en donde un estudiante puede culminar su participación en la entidad estudiantil ya sea por abandono voluntario, abandono involuntario o graduación del alumno (DesJardins et al., 2002, 2006; Lassibille y Navarro Gómez, 2008; Vallejos y Steel, 2017).
En el segundo caso, los tiempos son asumidos discretos, por el hecho de que los datos contruidos para el desarrollo de estos modelos son tomados por cada ciclo universitario, lo cual tiende a ser cada semestre.

2.1. Planteamiento del modelo

Dentro de los modelos de riesgos competitivos, es muy utilizado el método propuesto por Cox (1972) de odds proporcionales. Este modelo busca relacionar un conjunto de covariables al odds de un evento causa - específica con respecto al no evento en un determinado tiempo. Formalmente, el modelo puede ser presentado de la siguiente manera:

Sea T una variable aleatoria discreta que representa el tiempo, $T \in \{1, 2, ...\}$, sea $R \in \{1, 2, ..., \mathcal{R}\}$ una variable que representa el tipo de evento observado y sea $\mathbf{X}$ un vector de covariables. Definimos el riesgo causa-específica como:

$$\lambda(r, t|x) = \frac{P(R = r, T = t|\boldsymbol{X} = x)}{P(T \geq t|\boldsymbol{X} = x)}, \tag{2.1}$$

que es la probabilidad condicional de observar un evento de tipo r en el periodo t dado que ningún evento ocurrió antes, condicionado en $\boldsymbol{X} = x$.

Si especificamente definimos a $\boldsymbol{x_i} \in \mathbb{R}^k$ como el vector fila que contiene a los valores de las covariables para el i-ésimo individuo, $B = \left\{\beta_{(1)}, ...\beta_{(\mathcal{R})}\right\}$ con $\boldsymbol{\beta_{(r)}} \in \mathbb{R}^k$ como el vector de coeficientes de regresión para la causa específica r y $\delta = \{\delta_{11}, ..., \delta_{\mathcal{R}1}, \delta_{12}, ..., \delta_{\mathcal{R}2}, ...\}$ con δ_{rt} como los logaritmos de los odds basales (en adelante, log odds basales) de observar el evento r (con respecto al no evento) en el tiempo t, el modelo odd proporcional está definido por:

$$log\left\{\frac{\lambda(r, t|\delta, B; \boldsymbol{x_i})}{\lambda(0, t|\delta, B; \boldsymbol{x_i}}\right\} = \delta_{rt} + \boldsymbol{x_i}^\top \boldsymbol{\beta_{(r)}}, \qquad r = 1, ..., \mathcal{R},\ i = 1, ..., n,\ t = 1, 2, ... \tag{2.2}$$

donde

$$\lambda(0, t|\delta, B; \boldsymbol{x_i}) = 1 - \sum_{r=1}^{\mathcal{R}} \lambda(r, t|\delta, B; \boldsymbol{x_i}), \tag{2.3}$$

representa el riesgo de que no suceda ningun evento en el tiempo t. Equivalentemente (2.2) se puede reescribir como:

$$\lambda(r, t|\delta, B; \boldsymbol{x_i}) = \frac{\exp(\delta_{rt} + \boldsymbol{x_i}^\top \boldsymbol{\beta_{(r)}})}{1 + \sum_{r=1}^{\mathcal{R}} \exp(\delta_{rt} + \boldsymbol{x_i}^\top \boldsymbol{\beta_{(r)}})}. \tag{2.4}$$

El modelo (2.2) asume que cada individuo en la muestra no presenta ningún evento a través de sucesivos periodos de tiempo discretos hasta que experimenta cualquiera de los eventos de interés o es censurado por el fin del tiempo de observación. Entonces para cada individuo i, el evento de ocurrencia puede ser registrado usando una secuencia de variables dummy y_{itr} de la forma:

$$y_{itr} = \begin{cases} 0, & \text{si el individuo } i \text{ no experimenta el evento } r \text{ en el periodo } t, \\ 1, & \text{si el individuo } i \text{ experimenta el evento } r \text{ en el periodo } t. \end{cases} \tag{2.5}$$

El modelo asume que los eventos no se repiten, por lo que la secuencia de los valores de y_{itr} puede tomar uno de los dos valores para cada tiempo t, 0 en cada periodo de tiempo en que el individuo no experimento el evento y 1 si es que el individuo experimenta el evento por alguna de las causas especificas en alguno de los periodos observados.

Por ejemplo, si un individuo es censurado por la derecha este presentará un registro de la variable y_{itr} igual a cero para todos los tiempos observados, caso contrario, si un individuo no es censurado este presentará un registro de la variable y_{itr} igual a uno en el tiempo que se dio el evento por alguna de las causas específicas. La información sobre si un individuo es censurado puede ser denotado por C_i, donde sus valores c_i son definidos como:

$$c_i = \begin{cases} 0, & \text{si el individuo } i \text{ no es censurado,} \\ 1, & \text{si el individuo } i \text{ es censurado.} \end{cases} \tag{2.6}$$

Sea j_i el último periodo observado, siguiendo a Allison (1982) y Singer y Willet (1993) la construcción de la verosimilitud observada proviene de dos tipos de contribuciones: (a) de los individuos no censurados, como la probabilidad que ocurra el evento en el periodo j_i, y (b) de los individuos censurados, como la probabiidad que el evento ocurra después del periodo de tiempo j_i.

La probabilidad que un individuo experimente el evento r en el periodo de tiempo j_i puede ser escrito como un producto de términos, uno por periodo, describiendo la probabilidad condicional que el evento no ocurra en el periodo 1 hasta j_i -1 pero ocurra en el periodo j_i.

$$\begin{aligned} P(T_i = j_i, R_i = r|\boldsymbol{x_i}) &= P(T_i = j_i, R_i = r|T_i \geq j_i, \boldsymbol{x_i})P(T_i \neq j_i - 1|T_i \geq j_i - 1, \boldsymbol{x_i}) \times ... \\ &\quad \times P(T_i \neq 2|T_i \geq 2, \boldsymbol{x_i})P(T_i \neq 1|T_i \geq 1, \boldsymbol{x_i}) \\ &= P(T_i = j_i, R_i = r|T_i \geq j_i, \boldsymbol{x_i}) \prod_{j=1}^{j_i-1} P(T_i \neq j|T_i \geq j, \boldsymbol{x_i}). \end{aligned} \tag{2.7}$$

Reformulando (2.7) en terminos de λ_{rt} definido en (2.1), tenemos

$$P(T_i = j_i, R_i = r|\boldsymbol{x_i}) = \lambda(r, j_i|\boldsymbol{x_i}) \prod_{j=1}^{j_i-1} \left(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})\right). \tag{2.8}$$

Por otro lado, la probabilidad que un individuo censurado experimente el evento después del periodo j_i puede ser contruido similarmente como:

$$\begin{aligned} P(T_i > j_i|\boldsymbol{x_i}) &= P(T_i \neq j_i|T_i \geq j_i, \boldsymbol{x_i})P(T_i \neq j_i - 1|T_i \geq j_i - 1, \boldsymbol{x_i}) \times ... \\ &\quad \times P(T_i \neq 2|T_i \geq 2, \boldsymbol{x_i})P(T_i \neq 1|T_i \geq 1, \boldsymbol{x_i}) \\ &= \prod_{j=1}^{j_i} P(T_i \neq j|T_i \geq j, \boldsymbol{x_i}) \\ &= \prod_{j=1}^{j_i} \left(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})\right). \end{aligned} \tag{2.9}$$

Si asumimos que los individuos en la muestra son independientes, la función de verosimilitud es simplemente el producto de las probabilidades, $P(T_i = j_i)$ en el caso de los individuos donde se observan los eventos y $P(T_i > j_i)$ en el caso de los censurados.

$$L = \prod_{i=1}^{n} \left[P(T_i = j_i, R_i = r_i|\boldsymbol{x_i})\right]^{1-c_i} \left[P(T_i > j_i|\boldsymbol{x_i})\right]^{c_i}. \tag{2.10}$$

Reemplazando (2.8) y (2.9) en (2.10) tenemos:

$$\begin{aligned} L &= \prod_{i=1}^{n} \left[\lambda(r_i, j_i|\boldsymbol{x_i}) \prod_{j=1}^{j_i-1}(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i}))\right]^{1-c_i} \left[\prod_{j=1}^{j_i}(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i}))\right]^{c_i} \\ &= \prod_{i=1}^{n} \left[\lambda(r_i, j_i|\boldsymbol{x_i})^{1-c_i} \prod_{j=1}^{j_i-1}(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i}))^{1-c_i}\right] \left[\frac{\prod_{j=1}^{j_i}(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i}))}{\prod_{j=1}^{j_i}(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i}))^{1-c_i}}\right] \\ &= \prod_{i=1}^{n} \lambda(r_i, j_i|\boldsymbol{x_i})^{1-c_i} \left(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})\right)^{c_i} \prod_{j=1}^{j_i-1} \left(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})\right). \end{aligned} \tag{2.11}$$

Note que (2.11) puede ser modificada para introducir el indicador de historia del evento y_{ijr}. La idea es que si un individuo i es no censurado ($c_i = 0$), el evento r ocurre en el periodo de tiempo j_i; entonces el indicador y_{ijr} es igual a cero para todo los periodos de tiempo excepto para al último donde $j = j_i$, el cual el valor es 1. Por otro lado, si el individuo i es censurado ($c_i = 1$), el evento no ocurre en ningún periodo de tiempo, entonces el indicador y_{ijr} es igual a cero para todos los periodos. Entonces podemos escribir:

$$\begin{aligned} \prod_{j=1}^{j_i} \prod_{r=1}^{\mathcal{R}} \lambda(r_i, j|\boldsymbol{x_i})^{y_{ijr}} &= \begin{cases} \lambda(r_i, j_i|\boldsymbol{x_i}), & c_i = 0 \\ 1, & c_i = 1 \end{cases} \\ &= \lambda(r_i, j_i|\boldsymbol{x_i})^{1-c_i}. \end{aligned} \tag{2.12}$$

Por otro lado, si definimos y_{ij0} como:

$$y_{it0} = \begin{cases} 0, & \text{si el individuo } i \text{ no experimenta el no-evento en el periodo } t, \\ 1, & \text{si el individuo } i \text{ experimenta el no-evento en el periodo } t. \end{cases} \tag{2.13}$$

El último factor de (2.11) puede ser expresado como:

$$\begin{aligned} \prod_{j=1}^{j_i} \left(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})\right)^{y_{ij0}} &= \begin{cases} \prod_{j=1}^{j_i-1}(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})), & c_i = 0 \\ \prod_{j=1}^{j_i}(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})), & c_i = 1 \end{cases} \\ &= \left(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i}\right)^{c_i} \prod_{j=1}^{j_i-1} \left(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})\right). \end{aligned} \tag{2.14}$$

Entonces sustituyendo (2.14) en (2.11):

$$
\begin{aligned}
L &= \prod_{i=1}^{n}\prod_{j=1}^{j_i}\prod_{r=1}^{\mathcal{R}} \lambda(r_i, j|\boldsymbol{x_i})^{y_{ijr}} \left(1 - \prod_{r=1}^{\mathcal{R}} \lambda(r, j|\boldsymbol{x_i})\right)^{y_{ij0}} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{j_i}\prod_{r=1}^{\mathcal{R}} \lambda(r_i, j|\boldsymbol{x_i})^{I(y_{ij=r})}. \qquad (2.15)
\end{aligned}
$$

La ecuación (2.15) tiene la forma de una verosimilitud derivada de un modelo multinomial. Este modelo ha sido ampliamente estudiado en la literatura (Agresti, 2003).

2.2. Problema de la separación cuasi − perfecta

Si bien el problema de la estimación parece estar resuelta desde un punto de vista frecuentista, un problema particular del tipo de datos utilizados para el estudio de la deserción estudiantil puede conllevar a una imposibilidad de la estimación por este método.

Es natural pensar que los eventos en los que estamos interesados en estudiar para este caso puedan estar claramente separados por los rangos de los tiempos de supervivencia. Por ejemplo, el hecho de observar un evento de graduación implica que el estudiante debió pasar los requerimientos necesarios para poder realizar dicho evento, por lo que es natural pensar que este evento no se observe en los primeros semestres. Caso contrario, observar eventos de deserción puede ser más natural en los primeros semestre, dado que a medida que el estudiante esta más cerca a culminar este tiene menos incentivos de abandonar la universidad.

Esta separación de los eventos es conocido como una separación cuasi - completa, es decir, que los eventos observados son casi completamente separables por una o más covariables asociadas al evento, para nuestro caso el tiempo de supervivencia. Al no haber solapamiento de los eventos la estimación de los parámetros asociados tienden al infinito, causando problemas de convergencia o estimaciones incorrectas del parámetro (Albert y Anderson, 1984; Singer y Willet, 1993; Vallejos y Steel, 2017).

Los predictores asociados a captar los efectos en cada tiempo estan vinculados a los log odds basales δ_{rt}, por ejemplo si se define que las graduaciones no se pueden observar durante el segundo semestre de inscripción, la función de verosimilitud maximizará cuando el riesgo causa − específica relacionado con las graduaciones sea igual a cero en el instante $t = 2$, por lo que la estimación del parámetro δ_{r2} podra ser cercano a ∞.

Bajo el problema antes expuesto, una de las soluciones cotidianas utilizadas en la literatura relacionada a estos temas es utilizar un método bayesiano que permita incorporar información adicional al parámetro solucionando los problemas de estimación clásicos. Dicho método se expone en el siguiente capítulo.

Capítulo 3

Modelo de riesgos competitivos bayesiano

Una solución al problema de la separación cuasi-perfecta entre los eventos se puede implementar usando inferencia bayesiana, que permite adicionar información a través de un conocimiento a priori a los log odds basales δ_{rt}. En particular dentro de los métodos bayesianos existen diferentes formas para tratar estos problemas; sin embargo, algunas suelen complicarse al ser extendidas a modelos multinomiales (ej. la priori de Jefreys).

La idea clave detrás de agregar información a priori a la estimación de los parámetros es la poca probabilidad que en un cambio típico en una variable modifique en gran medida la probabilidad estimada de un evento, por ejemplo un cambio en 5 en la escala logística que correspondería a cambios en la probabilidad estimada del evento de 0,01 a 0,5. Estos cambios podrían mas bien ser originarios por los problemas de separación antes mencionados o problemas de colinealidad muy comunmente vistos en los análisis de regresión.

Algunos enfoques capturan esta idea proponiendo distribuciones a prioris específicas como el enfoque de medias condicionales de Bedrick et al. (1996) o el método de Witte et al. (1998) que asignan distribuciones a priori al caracterizar los efectos esperados en rangos debilmente informativos; sin embargo, en la medida que se busca una mayor generalización de la información brindada por una distribución a priori con respecto a los límites que pueden tomar los coeficientes, Gelman et al. (2008) proprone un método que permita el uso de una restricción genérica en lugar de informaciones específicas que permita mayor flexibilidad enfocadas en los coeficientes de regresión utilizando la distribución Cauchy como una opción conservadora de la familia t - Student.

3.1. Elección de las prioris del modelo

3.1.1. La priori Cauchy

Sea $X \sim Cauchy(x_0, \gamma)$, entonces su función de densidad de probabilidad (fdp) es dada por:

$$f(x; x_0, \gamma) = \frac{1}{\pi}\left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right], \qquad x \in \mathbb{R} \tag{3.1}$$

donde x_0 es el parámetro de localización y γ el parámetro de escala

Para el caso de deserción estudiantil, Vallejos y Steel (2017) utiliza el método propuesto por Gelman et al. (2008) el cual propone distribuciones a priori Cauchy debilmente informativas (media 0 y escala 0.25) para limitar el rango de los efectos reales del parámetro

sobre la estimación regularizando las inferencias extremas que se obtienen usando máxima verosimilitud.

Para nuestro caso práctico el problema de la estimación por máxima verosimilitud se enfocaba en la estimación de los log odds basales que capturaban el efecto del tiempo de supervivencia sobre los eventos. Estas estimaciones podrian resultar en valores muy altos en los semestres en que los eventos como la graduación no ocurrían.

Bajo la lógica de las prioris Cauchy lo que se busca es limitar el rango de la estimación de los log odds basales lejanos a ∞ obteniendo con ello asignar menor probabilidad a los ratios de riesgo causa - específica cercanas a cero para todos los periodos.

La elección de la distribución Cauchy dentro de la familia t - Student responde una opción más conservadora con respecto a la información a priori sobre el rango que podría tomar los coeficientes de regresión, asimismo, esto resulta ser más flexible, robusta y estable en la apliación a una regresión logística.

En la Figura 3.1 se muestra la fdp de esta distribución para diversos valores sus parámetros, por ejemplo, se observa que a mayor escala de la distribución tiende a dar colas más anchas, es decir, da mayor probabilidad a la ocurrencia de valores extremos.

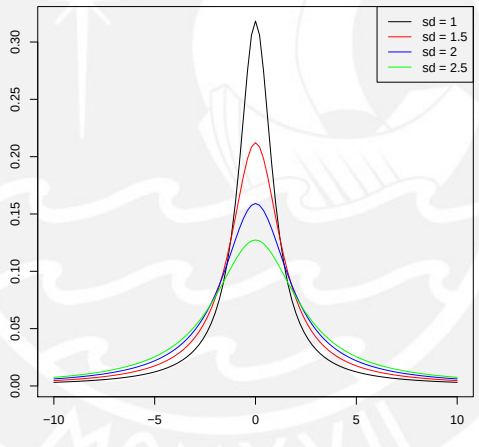


Figura 3.1: Función de densidad de la distribución Cauchy centrada en cero

El uso del modelo propuesto puede traer algunos inconvenientes, por el gran número de parámetros a estimar. Por ejemplo, si tenemos un total de tiempos τ tendríamos $\mathcal{R} \times \tau$ diferentes $\delta_{rt}{}'s$. Una solución al problema fue propuesto por Scott y Kennedy (2005) asignando una única causa-específica log-odds δ_{rt_0} al periodo $[t_0, \infty)$, fijando t_0 a un valor arbitrario en donde la mayoría de las personas ya experimentaron uno de los eventos (o censura).

Según lo antes dicho $\boldsymbol{\delta_r} = (\delta_{r1}, ..., \delta_{rt_0})^\top$ definido previamente en (2.2), por Vallejos y Steel (2017) tiene una distribución a priori:

$$\delta_r \sim Cauchy_{t_0}(\mathbf{0_{t_0}}, \omega^2 I_{t_0}), \qquad r = 1, ..., \mathcal{R},$$

donde I_{t_0} denota la matriz identidad de dimensión t_0, $\mathbf{0_{t_0}}$ es un vector de $0s$ y el ω a ser un indicador de precisión: pequeños valores de ω^2 resulta en prioris ajustadas, por otro lado,

valores grandes asignan mayor probabilidad a grandes valores negativos (y positivos) de los lod odds basales δ_{rt} (Figura 3.1).

Equivalentemente la distribución Multivariada Cauchy se puede escribir como:

$$\delta_r|\Lambda_r \quad \sim \quad N_{t_0}(\mathbf{0_{t_0}}, \Lambda_r^{-1}\omega^2 I_{t_0})$$

donde

$$\Lambda_r \quad \sim \quad \Gamma\left(\frac{1}{2}, \frac{1}{2}\right), \qquad r = 1, ..., \mathcal{R}. \tag{3.2}$$

Algunas propiedades de la Multivarada Cauchy pueden ser vistas en Bian y Dickey (1991).

3.1.2. Hyper g prioris generalizados

La inclusión de covariables al modelo bayesiano propuesto por Vallejos y Steel (2017) incorpora nuevos parámetros en la estimación bayesiana como los coeficientes de regresión. Bajo el enfoque propuesto por los autores, se propone el esquema de hyper g prioris generalizados propuesto por Sabanés y Held (2011) como una extensión de las g prioris introducidas por Zellner (1986) para los modelos lineales clásicos con distribución normal.

La idea detrás de las g prioris es no solo incorporar la incertidumbre con respecto a los parámetros del modelo, sino también sobre la composición del modelo. A diferencia de los modelos bayesianos de selección de variables que incorporan un parámetro de inclusión $\gamma \in \{0, 1\}^k$ para todas las k covariables disponibles, la idea aqui es considerar la foma del vector de covariables $\boldsymbol{x_{\gamma,i}}$ que incluye las diferentes transformaciones de las variables originales. Por ejemplo, cuando γ indica una transformación cuadratica de $\boldsymbol{x_i}$, entonces la forma del vector de covariables sigue $\boldsymbol{x_{\gamma,i}} = (x_i, x_i^2)^\top$ y entonces una priori sobre los coeficientes de regresión del modelo son requeridos $f(\beta_0, \beta_\gamma|\gamma)$ para todos los modelos $\gamma \in \Gamma$ donde Γ es el espacio de modelos posibles.
La verosimilitud a posteriori sigue de:

$$f(\gamma|y) \propto f(y|\gamma)f(\gamma), \qquad \gamma \in \Gamma. \tag{3.3}$$

$$f(y|\gamma) = \int_{\mathbb{R}^{p_\gamma+1}} f(y|\beta_0, \beta_\gamma, \gamma)f(\beta_0, \beta_\gamma|\gamma)d\beta_0 d\beta_\gamma. \tag{3.4}$$

Zellner (1986) propuso la g priori para el caso especial del modelo lineal normal clásico con varianza del error conocido ϕ y $w_i = 1$ como los pesos de la dispersión.

$$\beta_\gamma|g, \phi \sim N_{p_\gamma}(\mathbf{0_{p_\gamma}}, g\phi(\boldsymbol{X}_{\boldsymbol{\gamma}}^\top \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}), \tag{3.5}$$

donde $\boldsymbol{X_\gamma} = (x_{\gamma 1}, ..., x_{\gamma n})^\top$ es la matriz de diseño centrada.
El hiperparámetro $g > 0$ se puede interpretar como la inversa del tamaño relativo de la

muestra a priori por lo tanto su influencia es bastante fuerte. Valores mayores de g conducen a la preferencia de modelos menos complejos, fenómeno conocido como la paradoja de Lindley-Jefreys.

Liang et al. (2008) propuso el hyper g priori que es un caso especial de la incompleta - gamma inversa priori de Cui y George (2008) obteniendo una forma cerrada para la verosimilitud marginal $f(y|\gamma)$. La extensión de Sabanés y Held (2011) incorpora la generalización de las hyper g prioris para los modelos lineales generalizados,

$$\beta_\gamma|g,\phi \sim N_{p_\gamma}(\mathbf{0}_{\boldsymbol{p}_{\boldsymbol{\gamma}}}, g\phi c(\boldsymbol{X}_{\boldsymbol{\gamma}}^\top W \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}), \tag{3.6}$$

que incorpora la constante $c = v(h(0))dh/d\eta(0)^{-2}$ y la matriz de pesos $W = diag(w)$, asimismo en el caso logístico la g priori puede ser definido como:

$$\beta_\gamma|g,\sim N_{p_\gamma}(\mathbf{0}_{\boldsymbol{p}_{\boldsymbol{\gamma}}}, 4g(\boldsymbol{X}_{\boldsymbol{\gamma}}^\top \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}). \tag{3.7}$$

Finalmente para el caso concreto del modelo de riesgo competitivos las g priori quedan definidas como:

$$\beta_{(r)}|g_r,\sim N_k(\mathbf{0}_{\boldsymbol{k}}, 4g_r(\boldsymbol{X}^\top \boldsymbol{X})^{-1}), \qquad r = 1, ..., \mathcal{R}, \tag{3.8}$$

manteniendo las propiedades de ser invariantes a las transformaciones a escala de las covariables e incorporando la estructura de correlación entre ellas. Asimismo, dada la sensibilidad de la posteriori a los valores de las $g_1, ..., g_\mathcal{R}$ se incorpora la hyper priori g/n de Liang et al. (2008) de la forma:

$$\pi(g_r) = \frac{1}{n}\left(1 + \frac{g_r}{n}\right)^{-2}. \tag{3.9}$$

3.2. Construcción del algoritmo de Gibbs

La construcción de un modelo bayesiano para una regresión logística es un problema díficil de implementar a diferencia de la estimación bayesiana de un modelo probit, el cual se simplifica bajo el método de la variable latente propuesto por Albert y Chib (1993) para el muestreo de la posteriori.

Muchos trabajos han tratado de responder a la problemática utilizando el enfoque de aumento de datos (por ejemplo, Holmes y Held (2006), Früwirth-Schnatter et al. (2009), Gramacy y Polson (2012)) que resultan ser aproximados o mayormente complicados dado que implican múltiples capas de variables latentes. Polson et al. (2013) propone un nuevo algoritmo de aumento de datos para la regresión logística bayesiana apelando a una nueva familia de distribuciones Polya - Gamma.

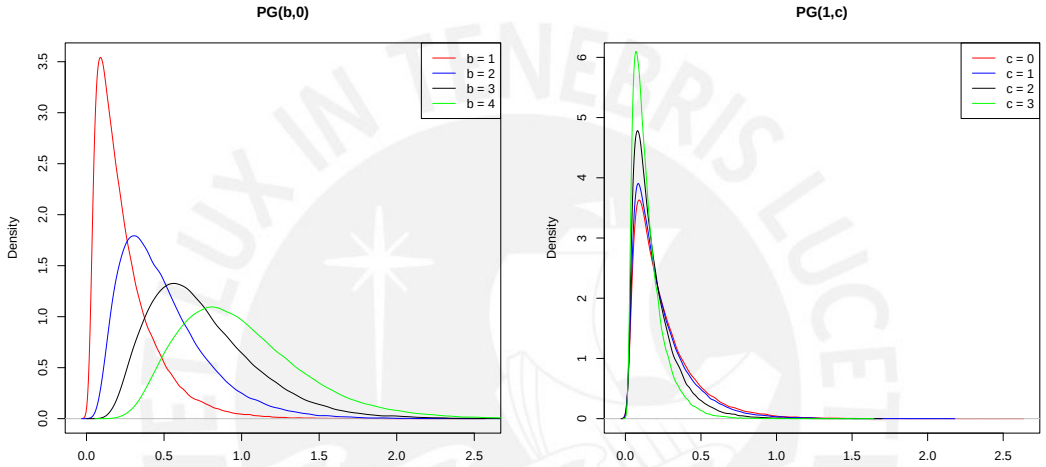
La idea principal detrás de este método es que las probabilidades binomiales parametrizadas por log odds pueden representarse como mezcla de una distribución normal con respecto

a una distribución de Polya - Gamma. Una variable X tiene una distribución Polya - Gamma (PG) con parámetros $b > 0$ y $c \in \mathbb{R}$, denotada como $X \sim PG(b, c)$, si

$$X \sim \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(\frac{k-1}{2})^2 + \frac{c^2}{(4\pi^2)}}, \tag{3.10}$$

donde $g_k \sim Gamma(b, 1)$ para $k \geq 1$.

Como se observa en el lado izquierdo de la Figura 3.2, cambios en el coeficiente b que denota los grados de libertad de las variables independientes gamma g_k modifican la forma de la distribución; por otro lado, como se observa en el lado derecho de la figura, cambios en el parámetro c controlan la escala de la distribución.



(a) Cambios grados de libertad (b) Cambios en la inclinación

Figura 3.2: Función de densidad Polya - Gamma

Entonces la identidad que permite que las probabilidades binomialies puedan ser representadas bajo una distribución Polya - Gamma es que para $b > 0$ se cumple que:

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b} e^{k\psi} \int_0^{\infty} e^{-\omega \frac{\psi^2}{2}} p(\omega) d\omega, \tag{3.11}$$

donde $k = a - \frac{b}{2}$ y $\omega \sim PG(b, 0)$. Cuando $\psi = \boldsymbol{X}^{\top}\boldsymbol{\beta}$ es función lineal de predictores, la integral es el kernel de una verosimilitud normal en $\boldsymbol{\beta}$. Asimismo la distribución condicional de ω, dado ψ, es también una distribución Polya - Gamma.

La demostración de esta identidad es importante pues de ello depende el algoritmo Gibbs utilizado para la estimación de los parámetos. Dicha demostración se puede observar en el apéndice A.1

La idea detrás del uso de la Polya - Gamma es construir un algoritmo de Gibbs simple para el modelo logístico Bayesiano. A diferencia del método propuesto por Albert y Chib (1993) para el modelo de regresión probit, la distribución a posteriori resultante es una mixtura normal de escala (donde sus componentes comparten una misma media) en vez de

una mixtura de nomal de locación (donde sus componentes comparten la misma varianza); por otro lado, las normales truncadas de Albert y Chib (1993) son reemplazadas por variables latentes Polya Gamma.

El algoritmo de Gibbs propuesto sigue de la siguiente lógica: Sea y_i el número de sucesos, n_i el número de intentos y $\boldsymbol{x_i} = (x_{i1}, ...x_{ip})^\top$ el vector de regresores de la observación $i \in 1, ..., N$. Sea $y_i \sim Binomial(n_i, 1/1 + e^{-\psi_i})$, donde $\psi_i = \boldsymbol{x_i}^\top \boldsymbol{\beta}$ son los log odds del evento. Finalmente, sea que $\boldsymbol{\beta}$ tenga una priori normal, $\beta \sim N(b, B)$, entonces:

$$\begin{aligned} (w_i|\beta) &\sim PG(n_i, \boldsymbol{x_i}^\top \boldsymbol{\beta}) && (3.12) \\ (\beta|y, \omega) &\sim Normal(m_\omega, V_\omega) && (3.13) \end{aligned}$$

donde

$$\begin{aligned} V_\omega &= (\boldsymbol{X}^\top \Omega \boldsymbol{X} + B^{-1})^{-1} && (3.14) \\ m_\omega &= V_\omega(\boldsymbol{X}^\top \boldsymbol{k} + B^{-1}\boldsymbol{b}), && (3.15) \end{aligned}$$

donde $k = (y_1 - \frac{n_1}{2}, ..., y_N - \frac{n_N}{2})$ y Ω es la matriz diagonal de $\omega_i's$.

Al igual que la identidad anterior, demostrar el algortimo de Gibbs para este ejemplo de una logística binomial es importante para luego introducir el algoritmo para nuestro caso de una logístico multinomial. En primer lugar, partimos que la verosimilitud de una logística binomial para una observación puede re-expresarse de la siguiente manera:

$$\begin{aligned} L_i(\beta) &= \left\{ \frac{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})} \right\}^{y_i} \left\{ 1 - \frac{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})} \right\}^{(1-y_i)} \\ &= \left\{ \frac{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})} \right\}^{y_i} \left\{ \frac{\frac{1}{1+\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}}{(\frac{1}{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})})^{y_i}} \right\} \\ &= \left\{ \frac{\frac{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}}{\frac{1}{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}} \right\}^{y_i} \frac{1}{1 + \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})} \\ &= \frac{\left\{\exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\right\}^{y_i}}{1 + \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})}. && (3.16) \end{aligned}$$

De (3.11) tenemos que la verosimilitud se puede expresar de la forma :

$$L_i(\beta) \propto \exp(k_i \boldsymbol{x_i}^\top \boldsymbol{\beta}) \int_0^\infty \exp\left\{ -\omega_i \frac{(\boldsymbol{x_i}^\top \boldsymbol{\beta})^2}{2} \right\} p(\omega|n_i, 0), \quad (3.17)$$

donde $k_i = y_i - \frac{n_i}{2}$; $p(\omega|n_i, 0) \sim PG(n_i, 0)$.

La condicional $p(\omega_i|n_i, \beta)$ puede expresarse como:

$$p(\omega|n_i,\beta) = \frac{e^{-\omega\frac{(\boldsymbol{x_i}^\top\boldsymbol{\beta})^2}{2}}p(\omega|n_i,0)}{\int_0^\infty e^{-\omega\frac{(\boldsymbol{x_i}^\top\boldsymbol{\beta})^2}{2}}p(\omega|n_i,0)d\omega}, \tag{3.18}$$

y de la densidad de la Polya - gamma podemos ver que $p(\omega|n_i,\beta) \sim PG(n_i, \boldsymbol{X}^\top\boldsymbol{\beta})$.
La distribución a posteriori de β queda expresado por:

$$\begin{aligned}
p(\beta|y,\omega) &\propto p(\beta)p(y|\omega,\beta) \\
&\propto p(\beta)\prod_{i=1}^{N} L_i(\beta|\omega_i) \\
&\propto p(\beta)\prod_{i=1}^{N}\exp(k_i\boldsymbol{x_i}^\top\boldsymbol{\beta})\int_0^\infty \exp\left\{-\omega_i\frac{(\boldsymbol{x_i}^\top\boldsymbol{\beta})^2}{2}\right\}p(\omega|n_i,0) \\
&\propto p(\beta)\prod_{i=1}^{N}\exp(k_i\boldsymbol{x_i}^\top\boldsymbol{\beta})\exp\left\{-\omega_i\frac{\boldsymbol{x_i}^\top\boldsymbol{\beta})^2}{2}\right\} \\
&\propto p(\beta)\prod_{i=1}^{N}\exp(k_i\boldsymbol{x_i}^\top\boldsymbol{\beta} - \omega_i\frac{(\boldsymbol{x_i}^\top\boldsymbol{\beta})^2}{2}) \\
&\propto p(\beta)\prod_{i=1}^{N}\exp\left\{\frac{\omega_i}{2}[-(\boldsymbol{x_i}^\top\boldsymbol{\beta})^2 + 2\frac{k_i\boldsymbol{x_i}^\top\boldsymbol{\beta}}{\omega_i}]\right\} \\
&\propto p(\beta)\prod_{i=1}^{N}\exp\left\{\frac{\omega_i}{2}(\boldsymbol{x_i}^\top\boldsymbol{\beta} - \frac{k_i}{w_i})^2\right\},
\end{aligned} \tag{3.19}$$

agregando en términos matriciales

$$p(\beta|y,\omega) \propto p(\beta)\exp\left\{-\frac{1}{2}(Z - X - \boldsymbol{\beta})^\top\Omega(Z - X\boldsymbol{\beta})\right\}, \tag{3.20}$$

donde $Z = \left\{\frac{k_1}{\omega_1}, \frac{k_2}{\omega_2}...\frac{k_n}{\omega_n}\right\}$ y $\Omega = diag(\omega_i)$
Si la priori de los coeficientes de regresión es normal ($\beta \sim Normal(b, B)$) tenemos:

$$\begin{aligned}
p(\beta|y,\omega) &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{b})^\top B^{-1}(\boldsymbol{\beta} - \boldsymbol{b})\right\}\exp\left\{-\frac{1}{2}(Z - X\boldsymbol{\beta})^\top\Omega(Z - X\boldsymbol{\beta})\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^\top B^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top B^{-1}\boldsymbol{b} + \boldsymbol{b}^\top B^{-1}\boldsymbol{b})\right\}\exp\left\{-\frac{1}{2}(Z^\top\Omega Z - 2Z^\top\Omega X\boldsymbol{\beta} + \boldsymbol{\beta}^\top X^\top\Omega X\boldsymbol{\beta})\right\} \\
&\propto \exp\left\{-\frac{1}{2}[\boldsymbol{\beta}^\top(B^{-1} + X^\top\Omega X)\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top(B^{-1}\boldsymbol{b} + X^\top\boldsymbol{k})]\right\} \\
&\propto Normal(\mu_m, \Sigma_m),
\end{aligned}$$

donde $\mu_m = \Sigma_m(B^{-1}\boldsymbol{b} + X^\top\boldsymbol{k})$ y $\Sigma_m = (B^{-1} + X^\top\Omega X)^{-1}$.

3.2.1. Distribución condicional del coeficiente de regresión

El método propuesto por Polson et al. (2013) visto arriba nos permite construir un muestreo de Gibbs para el modelo logístico multinomial, siguiendo las líneas de Holmes y Held (2006). Recordando el modelo logístico multinomial desarrollado en el apartado anterior, teníamos que y_{it} esta asociado a los eventos $0, 1, ..., \mathcal{R}$ con coeficientes de regresión $B^* = \beta^*_{(1)}, ...\beta^*_{(\mathcal{R})}$. De la verosimilitud escrita en el apartado anterior podemos derivar la verosimilitud condicional propuesto por Holmes y Held (2006):

$$L(\lambda) = \prod_{i=1}^{n} \prod_{j=1}^{j_i} \prod_{r=1}^{\mathcal{R}} [\lambda_{rij_i}]^{I(y_{ir}=r)}, \tag{3.21}$$

en terminos de $\boldsymbol{\beta}$:

$$L(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{j_i} \prod_{r=1}^{\mathcal{R}} \left[\frac{\exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}})}{1 + \sum_{r=1}^{\mathcal{R}} \exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}})} \right]^{I(y_{ir}=r)}, \tag{3.22}$$

si dividimos a los dos terminos por $1 + \sum_{r^* \neq r} \exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}})$, tenemos

$$L(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{j_i} \prod_{r=1}^{\mathcal{R}} \left[\frac{\exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})}{1 + \exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})} \right]^{I(y_{ir}=r)}, \tag{3.23}$$

donde $C_{ir} = log(1 + \sum_{r^* \neq r} \exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}}))$.
Podemos agregar la última productoria de los riesgos en competencia como si fuera una binomial.

$$L(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{j_i} \left[\frac{\exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})}{1 + \exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})} \right]^{I(y_{it}=r)} \left[1 - \frac{\exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})}{1 + \exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})} \right]^{I(y_{it}\neq r)}. \tag{3.24}$$

Operando y simplificando se obtiene la verosimilitud condicional de $\boldsymbol{\beta^*_{(r)}}$ condicionados a los valores fijos $\beta^*_{(1)}, ..., \beta^*_{(r-1)}, \beta^*_{(r+1)}..., \beta^*_{(\mathcal{R})}$ denotado por $\boldsymbol{\beta^*_{(-r)}}$

$$L(\beta_r|\beta_{-r}) = \prod_{i=1}^{n} \prod_{j=1}^{j_i} \frac{[\exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})]^{I(y_{it}=r)}}{1 + \exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})}. \tag{3.25}$$

De la misma manera que en la binomial, dada la forma de la función de verosimilitud construida, podemos hacer uso de la Polya Gamma para construir el algoritmo de Gibbs en el caso de un logístico multinomial. En concreto, la verosimilitud de la multinomial para una observación puede ser escrita como:

$$L_{itr}(\beta_r|\beta_{-r}) = \frac{[\exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})]^{I(y_{it}=r)}}{1+\exp(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})}. \tag{3.26}$$

La expresión anterior puede ser expresado en función de una Polya Gamma según el teorema antes mostrado:

$$L_{itr}(\beta_r|\beta_{-r}) \propto \exp[k_{itr}(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})] \int_0^\infty \exp\left\{-\eta_{itr}\frac{(\boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir})^2}{2}\right\} p(\eta|n_{itr},0), \tag{3.27}$$

donde $\eta_{itr} \sim PG(1,0)$, utilizando las condiciones de la condicional:

$$\eta_{itr}|B^* \sim PG(1, \boldsymbol{z_i}^\top \boldsymbol{\beta^*_{(r)}} - C_{ir}), \qquad t = 1,...,t_i; i = 1,...n. \tag{3.28}$$

La distribución a posteriori:

$$\beta^*_{(r)}, \beta^*_{(-r)}, y_{11},...,y_{nt_n} \propto P(\beta) \prod_{i=1}^{n}\prod_{j=1}^{j_i} L_{itr}(\beta_r|\beta_{-r}). \tag{3.29}$$

Asumiendo una distribución a priori normal multivariada de la forma $\beta^*_{(r)} \sim Normal_{t_0+k}(\mu_r, \Sigma_r)$ la distribución a posteriori queda de la forma:

$$\begin{aligned}
\beta^*_{(r)}, \beta^*_{(-r)}, y_{11},...,y_{nt_n} &\propto \exp\left\{-\frac{1}{2}[\boldsymbol{\beta^\top_{(r)}}(\Sigma_r^{-1} + Z^\top D_r Z)\boldsymbol{\beta_{(r)}} - 2\boldsymbol{\beta^\top_{(r)}}(\Sigma_r^{-1}\mu + Z^\top k_r)]\right\} \\
&\propto Normal_{t0+k}(m_r, V_r),
\end{aligned}$$

$$\begin{aligned}
V_r &= (\Sigma_r^{-1} + Z^\top D_r Z)^{-1}, \\
m_r &= V_r((\Sigma_r^{-1}\mu + Z^\top k_r), \\
k_r &= (k_{11r},...,k_{nt_nr})^\top, \\
k_{itr} &= I_{\{y_{it}=r\}} - \frac{1}{2} + \eta_{itr}C_{ir}, \\
\eta_r &= (\eta_{11r},...,\eta_{nt_nr})^\top, \\
D_r &= diag(\eta_r).
\end{aligned} \tag{3.30}$$

Para nuestro caso concreto las distribuciones a prioris tanto para los log odds basales y los coeficientes de regresión no son normales; sin embargo, como se mostró anteriormente estas podían ser llevados a normales incluyendo una transformación. Esta particularidad implica agregar un paso más al algoritmo de Gibbs planteado arriba.

Condicional completa del coeficiente de regresión

Al igual que el paso anterior partimos de la formulación de la conjunta agregando las funciones de densidad de las Hyper prioris:

$$y|. \propto p(\Lambda_R)p(\delta_{(r)}|\Lambda_r)p(g_r)p(\beta^*_{(r)}|g_r)\prod_{i=1}^{n}\prod_{j=1}^{j_i} L_{itr}(\beta_r|\beta_{-r}). \tag{3.31}$$

El producto independiente de los componentes de la Cauchy multivariada y las hyper - g prioris resulta en una distribución mixtura de normales. Dicho desarrollo es mostrado en Roberts y Rosenthal (2009) donde se muestra el paso del Metropolis - Hasting adaptativo.

Condicional completa del hyper - parámetro Λ

$$\Lambda_r|\delta_{(r)} \sim Gamma\left(\frac{t_0+1}{2}, \frac{\boldsymbol{\delta}^\top_{(\boldsymbol{r})}\boldsymbol{\delta}_{(\boldsymbol{r})}}{2\omega^2}\right), \qquad r = 1, ..., \mathcal{R}, \tag{3.32}$$

Condicional completa del hyper - parámetro g

$$g_r|\beta_{(r)} \propto {g_r}^{\frac{-k}{2}} \exp\left[-\frac{\boldsymbol{\beta}^\top_{(\boldsymbol{r})}X^\top X\boldsymbol{\beta}_{(\boldsymbol{r})}}{2g_r}\right]\pi(g_r) \qquad r = 1, ..., \mathcal{R}. \tag{3.33}$$

Algoritmo de Gibbs

En cada iteración y para cada $r \in \{1, ...\mathcal{R}\}$ muestreamos en primer lugar los log odds basales y los coeficientes de regresión a partir de una normal multivariada de la forma descrita en (3.30) tomando condiciones iniciales para los hyper - parámetros Λ_r y g_r que conforman la matriz de covarianzas Σ_r^{-1} de la distribución a priori, para la media de la distribución a priori y para η_{itr}. Posteriormente, muestreamos $\eta_{itr}|B*$ de la distribución Polya - Gamma de la forma descrita en (3.28). Seguidamente, muestreamos la hyper - parámetro Λ_r de la distribución gamma de la forma descrita en (3.32) con los valores actualizados de los coeficientes de regresión y los log odds basales. Finalmente, muestreamos el hyper - parámetro g_r aplicando Metropolis Hastings adaptativo siguiendo la distribución condicional completa de g_r mostrada en (3.33) con los valores muestreados en el paso previo de Λ_r, los log odds basales y los coeficientes de regresión.

$$\begin{aligned} &(1) && (\delta^\top_{(r)}, \beta^\top_{(r)})|\Lambda_r, g_r \\ &(2) && \Lambda_r|\beta_{(r)}, \delta_{(r)}, g_r \\ &(3) && g_r|\beta_{(r)}, \delta_{(r)}, \Lambda_{(r)}. \end{aligned} \tag{3.34}$$

Más detalles de las derivaciones de las condicionales completas se puede ver en Vallejos y Steel (2017).

Capítulo 4

Estudio de Simulación

Este capítulo presenta un estudio de simulación con el fin de evaluar los métodos explicados en el capítulo 3 utilizando los códigos en R desarrollados por Vallejos y Steel (2017) anexados a la tesis. En particular, se analiza si el modelo bayesiano expuesto permite recuperar los parámetros simulados y si logra corregir los problemas de sesgo resultante de los problemas de separación casi perfecta. Con dicho fin se evalúa criterios de sesgo porcentual medido como la diferencia porcentual entre la media de las simulaciones realizadas y el valor real del parámetro. Además para el análisis de la estimación bayesiana se presentan las cadenas de Markov y el gráfico de las densidades de las distribuciones a posteriori resultantes de la estimación.

4.1. Construcción de Escenarios

Se presentan dos escenarios, el primero que sigue el ejercicio de simulación presentado en Vallejos y Steel (2017) en el cual se genera la separación casi perfecta de los eventos 1 y 2 y el segundo escenario que incrementa el número de ocurrencias del evento 1 como un escenario más cercano al evento de graduación que se desarrolla en la aplicación.

Escenario A

Se realizan 10 000 réplicas para $n = 400$ valores simulados de los tres eventos obtenidos por los individuos en la muestra los cuales son generados de la aplicación de (2.4) en cada periodo observado para cada individuo. En la simulación se incluyeron tres covariables que fueron generadas por distribuciones binomiales con probabilidad $P = 0{,}5$ siendo una de ellas de orden 3. Asimismo, los parámetros correspondientes a los log odd basales para cada evento y periodo se muestran en las Tablas 4.1 y 4.2.Finalmente, los coeficientes de regresión de las variables simuladas se muestran en la Tabla 4.3 donde $X1 \sim bernoulli(0{,}5), X2 \sim binomial(2, 0{,}5)$ y $X3 \sim bernoulli(0{,}5)$.

Cuadro 4.1: Parámetros reales de los log odds basales Escenario A (A)

	Per. 1	Per. 2	Per. 3	Per. 4	Per. 5	Per. 6	Per. 7	Per. 8
δ_1	-8.23	-5.62	-4.82	-4.88	-5.96	-5.76	-4.89	4.37
δ_2	-8.51	6.07	5.42	4.89	4.81	4.30	4.46	4.96
δ_3	-3.14	1.31	0.28	0.77	-0.80	0.45	-2.22	0.64

Las condiciones antes mostradas son obtenidas del ejercicio de simulación realizado por Vallejos y Steel (2017) en el cual se genera el problema de la serparación casi perfecta para los dos primeros eventos, en donde el primero simula las graduaciones con eventos que se dan

Cuadro 4.2: Parámetros reales de los log odds basales Escenario A (B)

	Per. 9	Per. 10	Per. 11	Per. 12	Per. 13	Per. 14	Per. 15	Per. 16
δ_1	4.06	5.98	5.95	6.58	6.93	7.34	5.47	6.40
δ_2	4.37	3.54	-4.52	-4.92	-3.90	-4.35	-3.70	-4.48
δ_3	0.24	-0.86	-0.06	0.76	0.34	-0.14	1.85	-0.57

Cuadro 4.3: Parámetros reales de los coeficientes de regresión

	X1	X2.1	X2.2	X3
β_1	0	1	-1	0
β_2	1	0	0	0
β_3	0.5	-0.5	0	0

en su mayoría al final del tiempo de observación y el segundo simula la deserción involuntaria donde los evento ocurren en su mayoría al inicio del tiempo de observación.

En la Figura 4.1 se muestra las estimaciones no paramétricas de las funciones de riesgo de causa - específica promedio de las 10 000 simulaciones generadas bajo las condiciones mencionadas arriba. Como se observa en la función de riesgo del primer evento, los eventos comienzan a suceder a partir del semestre 10, por otro lado, la función de riesgo del segundo evento evidencia un mayor número de eventos en los primeros semestres y estos dejan de suceder a partir del semestre 10. Para el caso de la función de riesgo del tercer evento estos se dan a lo largo de los 16 semestres.

Escenario B

Las condiciones del segundo escenario son similares al primero con diferencia en los log odds basales utilizados para la simulación los cuales son elegidos de tal manera que conserven los problemas de separación casi - perfecta incluyendo una mayor cantidad de eventos del escenario que simula la graduación como un escenario más cercana a la realidad. Los parámetros correspondientes a los log odds basales para cada evento y periodo se muestran en los cuadros 4.4 y 4.5

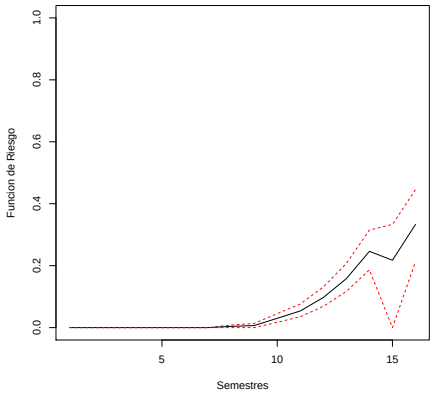
Cuadro 4.4: Parámetros reales de los log odds basales Escenario B (A)

	Per. 1	Per. 2	Per. 3	Per. 4	Per. 5	Per. 6	Per. 7	Per. 8
δ_1	-6.23	-5.62	-4.82	-4.88	-5.96	-5.76	-4.89	4.37
δ_2	-9.51	6.07	5.42	4.89	4.81	4.30	4.46	4.96
δ_3	-5.14	1.31	0.28	0.77	-0.80	0.45	-2.22	0.64

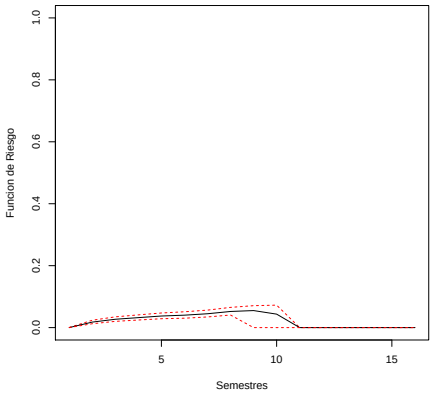
La Figura 4.2 se muestra las estimaciones no parámetricas de los ratios de riesgo causa - específica promedio de las 1 000 simulaciones generados bajo el nuevo escenario. La función de riesgo de causa - específica del primer evento confirma el escenario planteado donde este es el más material en eventos alrededor del 80 % de los eventos.

4.2. Estimación Clásica

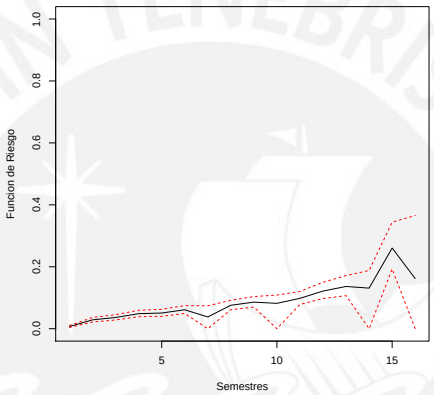
En base a estas especificaciones se estimó para cada simulación de datos el modelo de riesgos competitivos clásico usando el método de máxima verosimilitud mediante la función **multinom** del paquete **nnet** en R para los dos escenarios presentados.



(a) Evento 1



(b) Evento 2



(c) Evento 3

Figura 4.1: Estimación no parámetrica de las funciones de riesgo de causa - específica e intervalos de 95 % confianza: Escenario A.

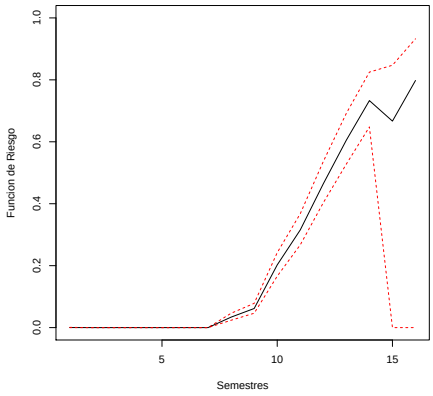
Cuadro 4.5: Parámetros reales de los log odds basales Escenario B (B)

	Per. 9	Per. 10	Per.11	Per. 12	Per. 13	Per. 14	Per. 15	Per. 16
δ_1	4.06	5.98	5.95	6.58	6.93	7.34	5.47	6.40
δ_2	4.37	3.54	-4.52	-4.92	-3.90	-4.35	-3.70	-4.48
δ_3	0.24	-0.86	-0.06	0.76	0.34	-0.14	1.85	-0.57

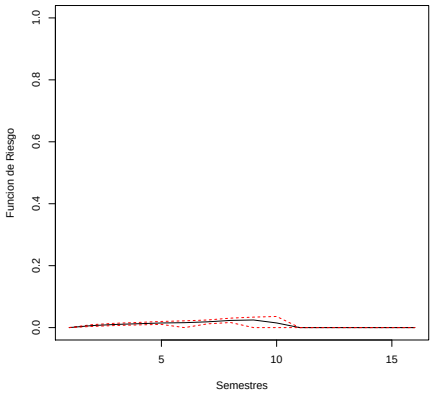
Escenario A

El resumen del resultado de la simulación bajo el escenario A se observa en la Tabla 4.6 donde se muestra el valor de los parámetros reales con los cuales se genero la simulación, la media de los valores estimados par las 10 000 simulaciones y el sesgo resultante para cada parámetro.

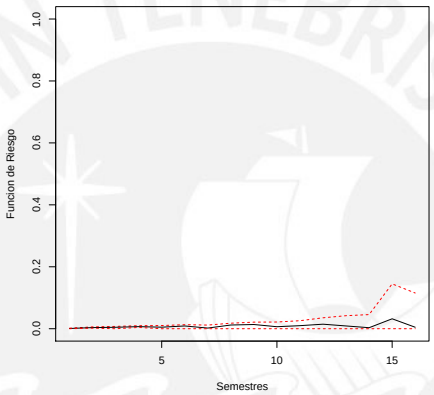
Asimismo, como caso particular de la Tabla 4.6, en la Figura 4.3 se muestra los histogramas de los log odds basales de experimentar el evento en el periodo 1 estimados bajo las 10 000 bases de datos generadas con el valor real utilizado (línea vertical). Como se observa en la figura, estas estimaciones resultan ser sesgadas para los log odds basales de los eventos 1 y 2, por otro lado el log odds basal del evento 3 resulta ser insesgado.



(a) Evento 1



(b) Evento 2



(c) Evento 3

Figura 4.2: Estimación no parámetrica de las funciones de riesgo de causa - específica e intervalos de 95 % confianza: Escenario B.

Por otro lado la Figura 4.4 muestra los histogramas del coeficiente de regresión β_1 bajo las 10 000 bases de datos generadas. Como se observa en la figura, estas estimaciones resultan ser insesgadas para los 3 eventos, lo cual confirma la intuición detrás del problema de separación casi perfecta.

Escenario B

El resumen del resultado de la simulación bajo el escenario B se observa en la Tabla 4.7 donde se muestra el valor de los parámetros reales con los cuales se genero la simulación, la media de los valores estimados par las 1 000 simulaciones y el sesgo resultante para cada parámetro.

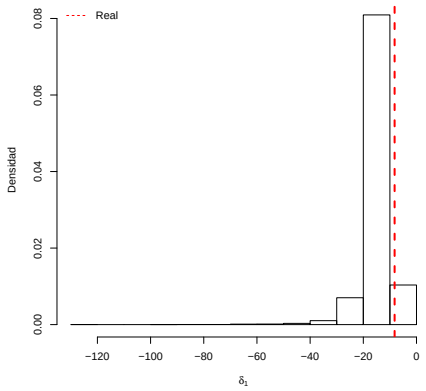
Asimismo, como caso particular de la Tabla 4.7, en la Figura 4.5 se muestra los histogramas de los log odds basales de experimentar el evento en el periodo 1 y del coeficiente de regresión de la primera covariable estimados bajo las 1 000 bases de datos generadas. Como se observa en la figura, estas estimaciones resultan ser igualmente sesgadas para los log odds basales e insesgadas para los betas, lo cual confirma la intuición detrás del problema de separación casi perfecta..

Cuadro 4.6: Resumen Escenario A: Clásico

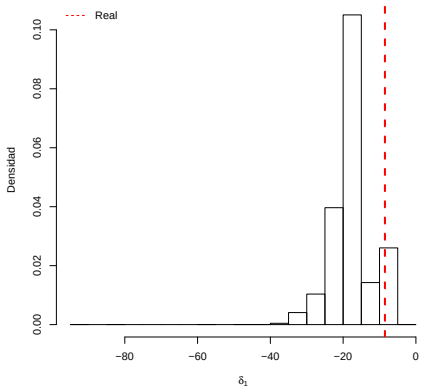
	Real 1	Real 2	Real 3	Sesgo 1	Sesgo 2	Sesgo 3	Per 1	Per 2	Per 3
δ_1	-8.23	-8.51	-3.14	8.17	9.22	0.04	99.24	108.33	1.24
δ_2	-5.62	6.07	1.31	-2.82	-9.17	-0.02	50.27	151.08	1.44
δ_3	-4.82	5.42	0.28	-2.87	-9.16	0.00	59.49	168.94	0.79
δ_4	-4.88	4.89	0.77	-3.19	-9.13	-0.01	65.34	185.64	-1.34
δ_5	-5.96	4.81	-0.8	-4.35	-9.13	0.26	73.02	189.72	31.94
δ_6	-5.76	4.3	0.45	-4.19	-8.90	0.00	72.72	206.89	0.44
δ_7	-4.89	4.46	-2.22	-3.22	-9.00	5.65	65.93	201.82	254.45
δ_8	4.37	4.96	0.64	-7.65	-9.11	-0.01	174.96	183.69	0.92
δ_9	4.06	4.37	0.24	-6.85	-8.65	0.03	168.82	197.97	11.59
δ_{10}	5.98	3.54	-0.86	-8.1	-5.64	1.63	135.5	159.22	189.32
δ_{11}	5.95	-4.52	-0.06	-8.11	-2.39	0.25	136.25	52.77	422.99
δ_{12}	6.58	-4.92	0.76	-8.13	-2.41	0.02	123.62	48.91	2.41
δ_{13}	6.93	-3.9	0.34	-8.15	-0.10	0.35	117.65	2.60	101.57
δ_{14}	7.34	-4.35	-0.14	-8.17	1.04	3.25	111.29	24.02	2320.10
δ_{15}	5.47	-3.7	1.85	-5.34	3.50	-0.01	97.63	94.49	0.74
δ_{16}	6.4	-4.48	-0.57	-7.67	4.95	13.86	119.91	110.53	2432.04
β_1	0	1	0.5	0.00	-0.03	-0.01			
β_2	1	0	-0.5	-0.04	-0.01	0.01			
β_3	-1	0	0	0.04	0.00	0.00			
β_4	0	0	0	0.01	0.00	0.00			

Cuadro 4.7: Resumen Escenario B: Clásico

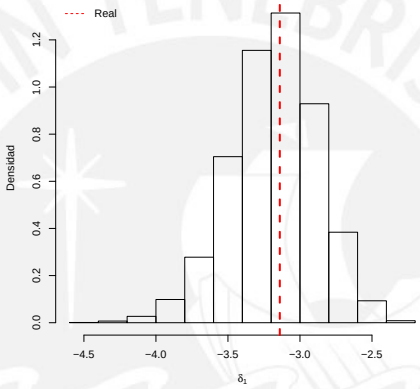
	Real 1	Real 2	Real 3	Sesgo 1	Sesgo 2	Sesgo 3	Per 1	Per 2	Per 3
δ_1	-6.23	-9.51	-5.14	2.53	8.25	0.93	-40.62	-86.76	-18.09
δ_2	-5.62	6.07	1.31	9.89	-8.16	-0.79	175.89	134.43	60.44
δ_3	-4.82	5.42	0.28	9.43	-8.13	-0.20	195.65	149.98	70.95
δ_4	-4.88	4.89	0.77	8.85	-8.09	-0.68	181.34	165.35	88.15
δ_5	-5.96	4.81	-0.8	7.31	-7.98	3.38	122.63	165.86	422.56
δ_6	-5.76	4.3	0.45	7.29	-7.06	-0.29	126.60	164.11	63.52
δ_7	-4.89	4.46	-2.22	7.88	-7.63	8.08	161.21	171.18	363.98
δ_8	4.37	4.96	0.64	-2.50	-8.01	-0.37	57.31	161.40	57.91
δ_9	4.06	4.37	0.24	-2.50	-6.67	0.35	61.65	152.57	143.81
δ_{10}	5.98	3.54	-0.86	-2.53	1.32	5.01	42.29	37.38	582.98
δ_{11}	5.95	-4.52	-0.06	-2.54	2.01	3.31	42.65	44.51	5524.25
δ_{12}	6.58	-4.92	0.76	-2.55	-0.26	4.47	38.77	5.23	587.98
δ_{13}	6.93	-3.9	0.34	-2.55	-0.13	16.14	36.86	3.39	4747.09
δ_{14}	7.34	-4.35	-0.14	-2.57	-0.82	25.19	34.95	18.94	17991.97
δ_{15}	5.47	-3.7	1.85	0.14	0.38	21.91	2.54	10.20	1184.35
δ_{16}	6.4	-4.48	-0.57	-1.34	-1.08	28.56	20.86	24.06	5009.82
β_1	0	1	0.5	-0.01	-0.07	-0.03			
β_2	1	0	-0.5	-0.03	0.01	0.24			
β_3	-1	0	0	0.03	0.01	0.00			
β_4	0	0	0	0.00	0.01	0.01			



(a) Evento 1



(b) Evento 2



(c) Evento 3

Figura 4.3: Histogramas de las 10 000 simulaciones del log odds basal del primer periodo para los tres eventos: Escenario A.

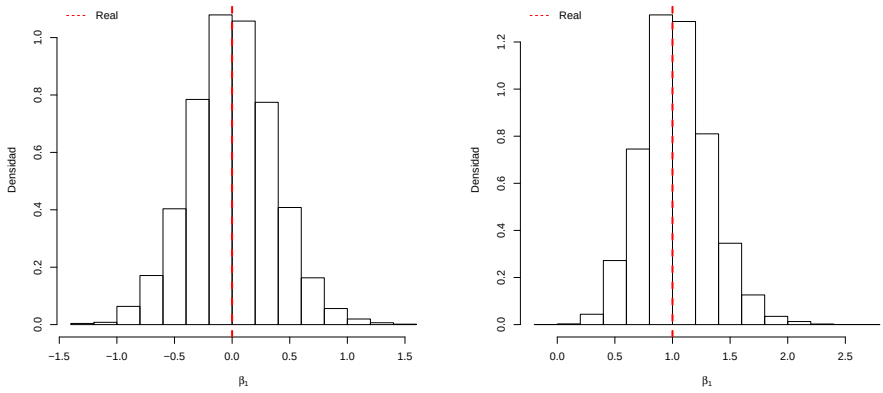
4.2.1. Estimación Bayesiana

El problema de estimación por el problema de la serpación de eventos se ve resuelto con la implementación de un modelo bayesiano que limita el valor de los logg odds basales en base a un conocimiento previo sobre un rango razonable de variación para los modelos logísticos. El modelo se estima bajo el algoritmo de Gibbs descrito en la ecuación 3.34, realizando 100 000 simulaciones de la distribución a posteriori con saltos de 10 en 10 para evitar la correlación entre los valores simulados y quemando las primeras 4 000 observaciones para mostrar la convergencia.

Escenario A

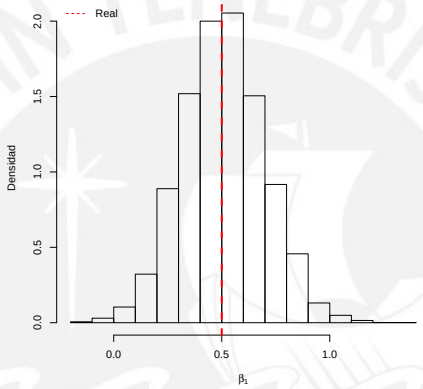
El resumen del resultado de la simulación bajo el escenario A se observa en la Tabla 4.8 donde se muestra el valor de los parámetros reales con los cuales se genero la simulación, la media de los valores estimados par las 10 000 simulaciones de la distribución a posteriori y el sesgo resultante para cada parámetro.

Asimismo, de manera particular se observa en la Figura 4.6 se observa las cadenas de Markov y las densidades para los log odds basales de experimentar el evento en el periodo



(a) Evento 1

(b) Evento 2



(c) Evento 3

Figura 4.4: Histogramas de las 10 000 simulaciones del coeficiente de la primera covariable para los tres eventos: Escenario A.

1 para una muestra representativa de la lógica descrita anteriormente. Las cadenas y las densidades muestran ser robustas y bien comportadas y los valores reales dentro del intervalo de confiabilidad para los tres eventos, por lo que el modelo resulta ser más robusto que lo observado en el modelo clásico.

Escenario B

El resumen del resultado de la simulación bajo el escenario B se observa en la Tabla 4.9 donde se muestra el valor de los parámetros reales con los cuales se genero la simulación, la media de los valores estimados par las 10 000 simulaciones de la distribución a posteriori y el sesgo resultante para cada parámetro.

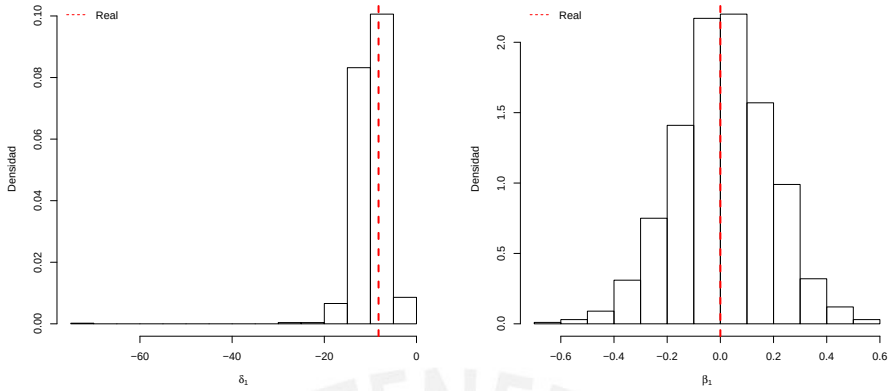
Asimismo, de manera particular se observa en la Figura 4.7 se observa las cadenas de Markov y las densidades para los log odds basales de experimentar el evento en el periodo 1 para una muestra representativa de la lógica descrita anteriormente. Las cadenas y las densidades muestran ser robustas y bien comportadas y los valores reales dentro del intervalo de confiabilidad para los tres eventos por lo que el modelo resulta ser más robusto que lo observado en el modelo clásico.

Cuadro 4.8: Resumen Escenario A: Bayesiano

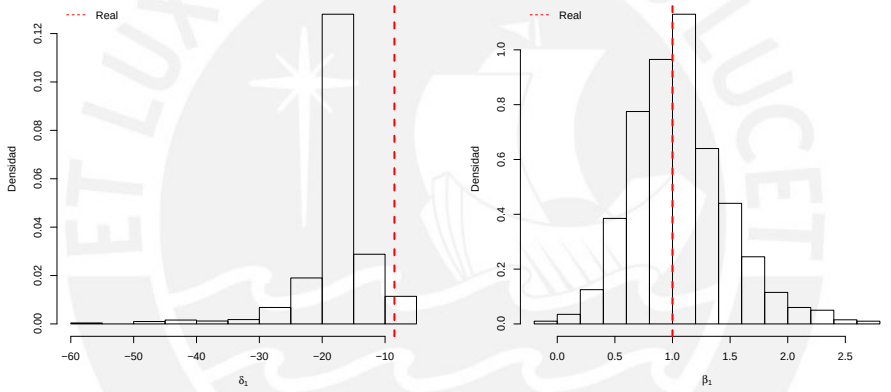
	Real 1	Real 2	Real 3	Sesgo 1	Sesgo 2	Sesgo 3	Per 1	Per 2	Per 3
δ_1	-8.23	-8.51	-3.14	-0.71	-1.94	-0.12	8.67	22.78	3.97
δ_2	-5.62	6.07	1.31	-0.88	1.68	0.01	15.66	27.73	0.69
δ_3	-4.82	5.42	0.28	-0.56	1.82	0.24	11.61	33.51	84.03
δ_4	-4.88	4.89	0.77	-0.46	1.96	0.06	9.45	40.11	8.02
δ_5	-5.96	4.81	-0.8	-2.33	1.21	0.49	39.06	25.24	60.94
δ_6	-5.76	4.3	0.45	-1.23	0.83	0.00	21.32	19.23	0.15
δ_7	-4.89	4.46	-2.22	-0.76	1.82	0.04	15.53	40.72	1.95
δ_8	4.37	4.96	0.64	1.71	1.67	0.5	39.06	33.70	78.86
δ_9	4.06	4.37	0.24	1.32	2.11	0.65	32.55	48.35	270.40
δ_{10}	5.98	3.54	-0.86	0.68	2.56	0.36	11.315	72.41	41.82
δ_{11}	5.95	-4.52	-0.06	0.48	-1.73	-0.23	7.99	38.37	386.26
δ_{12}	6.58	-4.92	0.76	0.43	-2.22	-0.56	6.54	45.05	73.51
δ_{13}	6.93	-3.9	0.34	1.50	-1.45	0.31	21.65	37.18	89.83
δ_{14}	7.34	-4.35	-0.14	0.96	-1.88	2.75	13.14	43.14	1964.41
δ_{15}	5.47	-3.7	1.85	1.09	-1.28	0.56	19.87	34.48	30.00
δ_{16}	6.4	-4.48	-0.57	0.78	-2.29	-0.83	12.16	51.16	145.86
β_1	0	1	0.5	-0.03	0.54	0.25			
β_2	1	0	-0.5	-0.49	-0.21	-0.06			
β_3	-1	0	0	-0.11	0.03	-0.08			
β_4	0	0	0	-0.15	-0.01	0.15			

Cuadro 4.9: Resumen Escenario B: Bayesiano

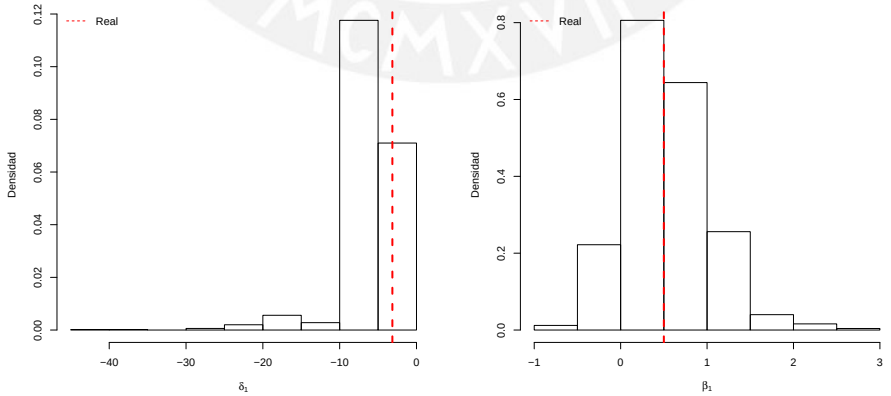
	Real 1	Real 2	Real 3	Sesgo 1	Sesgo 2	Sesgo 3	Per 1	Per 2	Per 3
δ_1	-6.23	-9.51	-5.14	-1.42	-2.36	0.45	22.72	24.77	8.79
δ_2	-5.62	6.07	1.31	-0.58	1.97	-0.56	10.24	32.5	42.64
δ_3	-4.82	5.42	0.28	1.30	3.14	-0.88	27.06	57.93	314.62
δ_4	-4.88	4.89	0.77	0.82	2.31	0.83	16.82	47.29	107.92
δ_5	-5.96	4.81	-0.8	-0.25	2.01	2.42	4.17	41.85	302.06
δ_6	-5.76	4.3	0.45	-0.03	2.76	-0.31	0.52	64.14	69.86
δ_7	-4.89	4.46	-2.22	0.61	3.91	-2.21	12.38	87.78	99.77
δ_8	4.37	4.96	0.64	1.36	1.97	-0.56	31.17	39.76	88.28
δ_9	4.06	4.37	0.24	1.69	1.08	-1.09	41.55	24.62	454.71
δ_{10}	5.98	3.54	-0.86	1.48	6.36	-0.67	24.68	179.58	78.13
δ_{11}	5.95	-4.52	-0.06	1.52	-1.79	-0.28	25.49	39.59	462.19
δ_{12}	6.58	-4.92	0.76	2.03	-2.42	0.12	30.91	49.23	15.81
δ_{13}	6.93	-3.9	0.34	1.76	-1.88	2.39	25.33	48.28	702.54
δ_{14}	7.34	-4.35	-0.14	2.05	-2.59	1.67	28.05	59.43	1189.34
δ_{15}	5.47	-3.7	1.85	1.21	-2.12	3.37	22.21	57.38	182.34
δ_{16}	6.4	-4.48	-0.57	2.13	-3.02	0.83	33.25	67.35	145.49
β_1	0	1	0.5	-0.02	0.19	-0.39			
β_2	1	0	-0.5	0.33	0.26	-0.56			
β_3	-1	0	0	-0.17	0.1	- 0.18			
β_4	0	0	0	0.06	0.02	0.01			



(a) Evento 1 (b) Evento 1

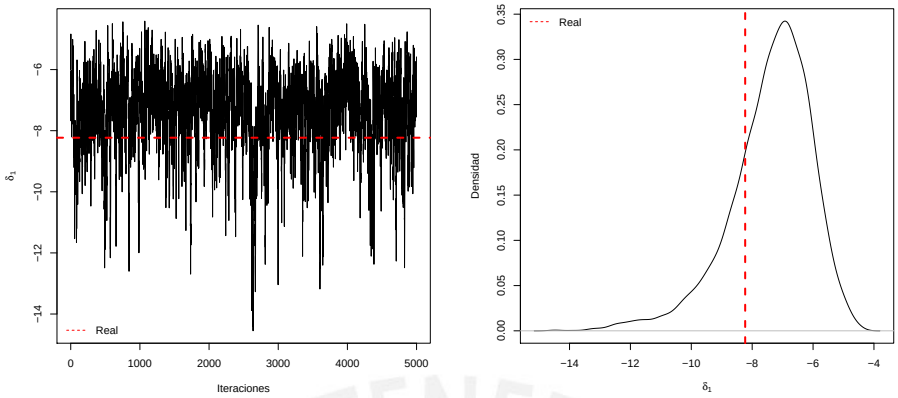


(c) Evento 2 (d) Evento 2

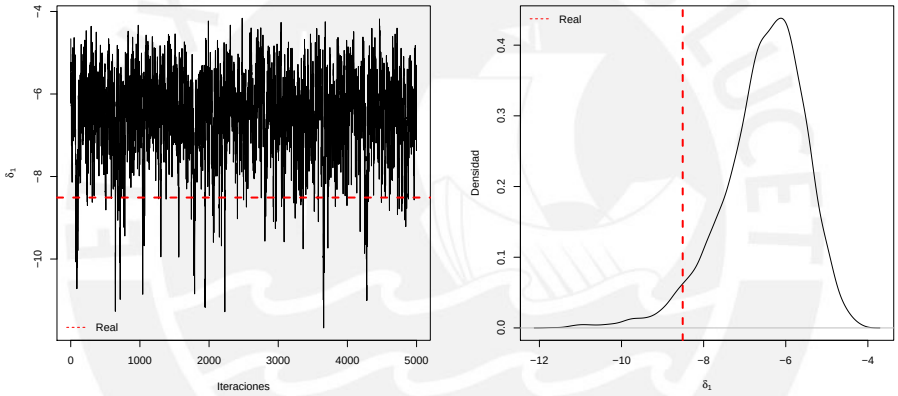


(e) Evento 3 (f) Evento 3

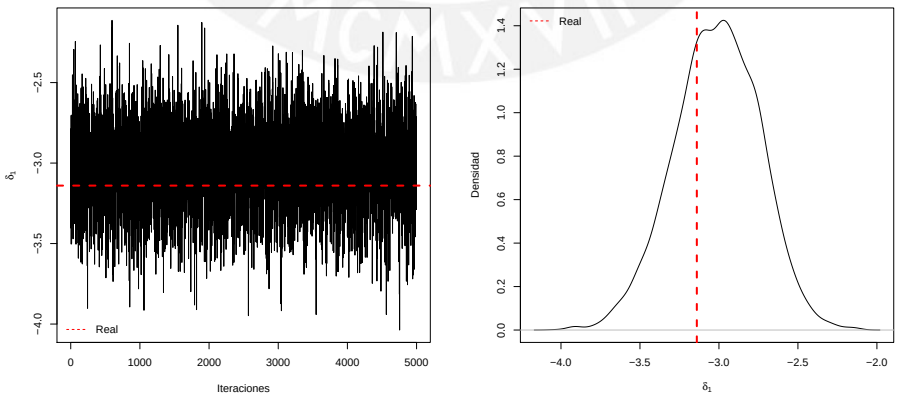
Figura 4.5: Histogramas de las 1000 simulaciones del log odds basal del primer periodo y del coeficiente de la primera covariable para los tres eventos: Escenario B



(a) Evento 1 (b) Evento 1

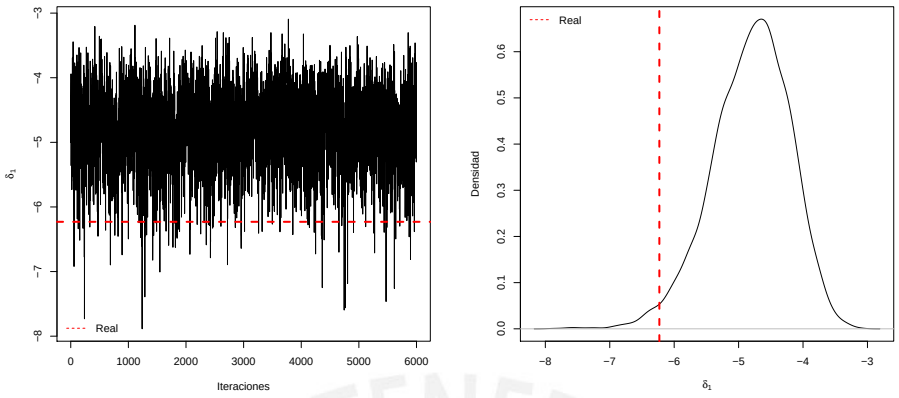


(c) Evento 2 (d) Evento 2

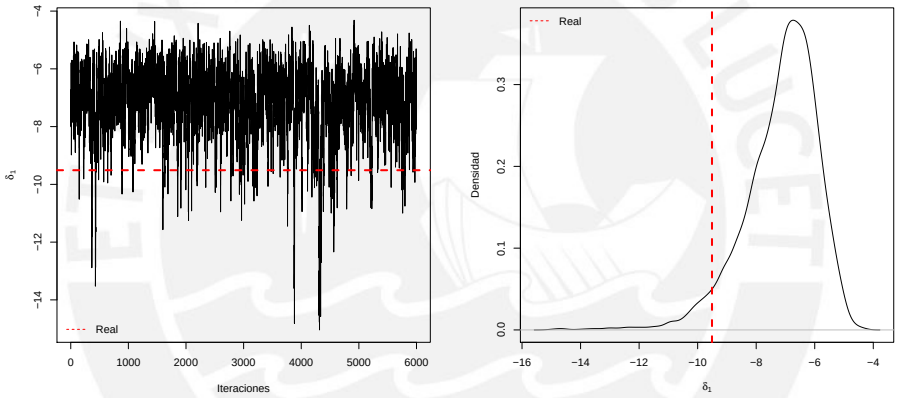


(e) Evento 3 (f) Evento 3

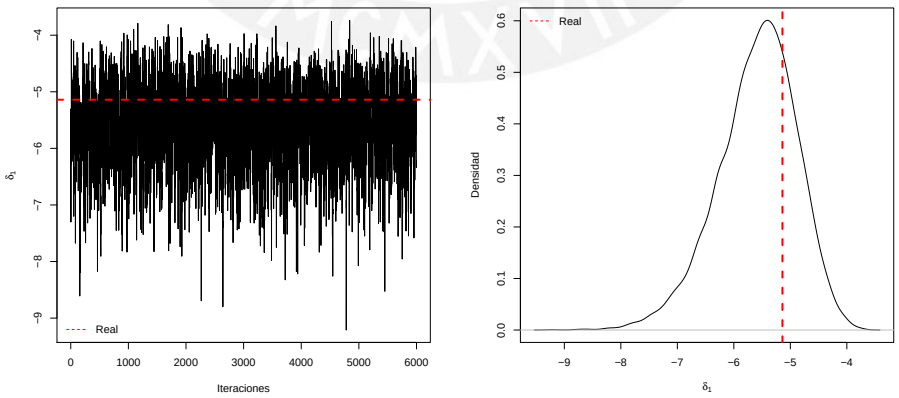
Figura 4.6: Cadena de Markov y densidades de la distribución a posteriori para el log odd basal del primer periodo para los tres eventos: Escenario A.



(a) Evento 1 (b) Evento 1



(c) Evento 2 (d) Evento 2



(e) Evento 3 (f) Evento 3

Figura 4.7: Cadena de Markov y densidades de la distribución a posterior para el log odd basal del primer periodo para los tres eventos: Escenario B.

Capítulo 5

Aplicaciones

5.0.1. Marco del estudio entre el 2004 y 2012

La aplicación del modelo mostrado se realliza a una muestra de la Pontificia Universidad Católica del Perú conformada por un total de 26 586 alumnos que ingresaron a la universidad utilizando los códigos en R desarrollados por Vallejos y Steel (2017) anexados a la tesis. El tiempo de observación de los alumos fue de un total de 18 semestres excluyendo los ciclos de verano.

Los datos estan conformados por cuatro grandes bloques de información correspondiente al momento de ingreso, a los ciclos matriculados, al ciclo de egreso y al resultado de cada materia cursada que engloban un total de 62 variables incluyendo variables que cambian en el tiempo. Asimsimo, las especialidades de los alumnos en la muestra conforman las unidades de artes, arquitectura y urbanismo, humanidades y ciencias sociales, ciencias formales e ingeniería; o educación. El Cuadro 5.1 muestra la descripción de los cuatro bloques, las observaciones, el número de variables, el total de alumnos y la llave de cada tabla.

Cuadro 5.1: Bases de datos

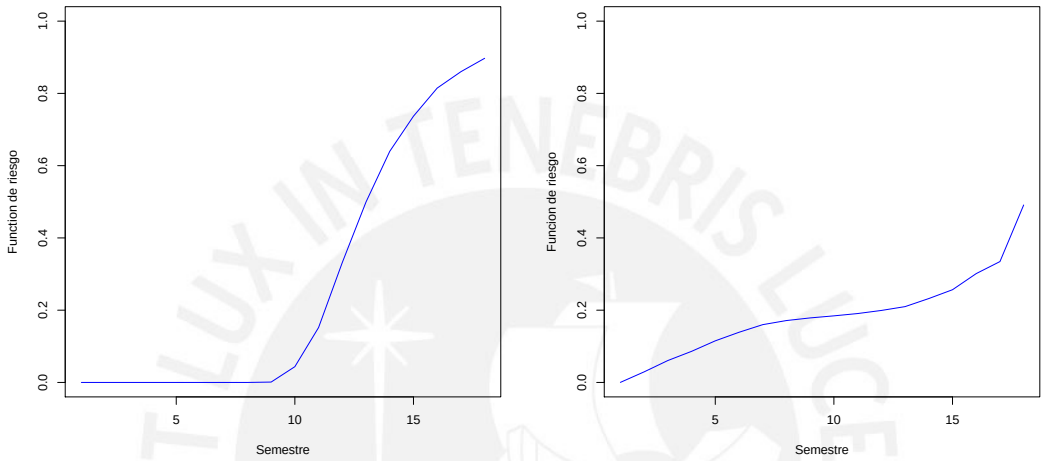
Tabla	Descripcion	Obs.	Var.	Alum.	Llave
Alumno - final	Variables al momento del ingreso	31 892	23	31 892	numsec
Alumno - ciclo matriculado	Variables por ciclo matriculado	276 971	21	31 892	numsec, cicmat
Alumno - ciclo egreso	Variables al momento del egreso	7 259	7	7 255	numsec
Alumno - cicmat- curso	Variable a nivel curso	1 289 614	11	31 892	numsec, cicmat, clavecur

Los eventos estudiados conforman la graduación y el abandono involuntario del alumno, siendo el primero la situación en que el alumno culmina la totalidad de los cursos requeridos por su especialidad, considerando únicamente el primer curso en que el alumno se graduo para los casos de alumnos que llevaban en paralelo más de una especialidad. Asimismo, se imputaron algunos eventos de graduación en los casos de alumnos que habian culminado la totalidad de los cursos en un periodo pero que no tuvieron la marca de egreso meses después posiblemente por un tema de pagos.

Con respecto al evento de abandono, la marca se definio por la variable **ielimciclo** que

señala si la culminación del ciclo conllevo a la eliminación del alumno. Para este evento se consideró como abandono si el alumno en el último ciclo observado mantenía la marca de eliminado, siendo que los alumnos que abandonaron una especialidad para iniciar otra no fueran considerados por el evento cuestión salvo si al culminar la fecha de observación este presentó la marca de eliminado. Esta definición se consideró debido a que se presentaba casos de alumnos con la marca eliminado pero que llegaron a graduarse en otra especialidad.

La Figura 5.1 presenta la estimación no paramétrica de la función de riesgo asociada a cada evento para los ciclos observados en la muestra.



(a) Función de riesgo asociado al evento de graduación

(b) Función de riesgo asociado al evento de abandono

Figura 5.1: Estimación no paramétrica de la función de riesgo por Kaplan y Meier

5.0.2. Estructura de la data

Las covariables consideradas en el estudio consideran dos niveles que consideran las características de los alumnos al momento del ingreso obtenida através de su declaración jurada y características pertenencientes a los ciclo matriculados por los alumnos que rescatan una temporalidad en la obervación del alumno.

Dentro del primer grupo consideramos variables como el sexo del alumno, si la madre o el padre del estan vivos, el grado de instrucción del padre y de la madre, si el alumno se encuentra trabajando o no, la modalidad de ingreso del alumno, area de primera matricula, tipo de colegio del alumno, departamento del colegio de origen y su ingreso familiar.

Cabe resaltar que se paso por prosecamiento de datos que consideró la imputación de missing values con la moda en caso se tratara de una variable categórica y la media en caso de variables continuas. La materialidad de los datos imputados con respecto al total de datos observados fue marginal.

5.0.3. Resultados

De los 25 586 estudiantes, 4 084 (15,36 %) presentaron el evento de graduación, mientras que 3 764 (14,16 %) presentaron el evento de abandono involuntario. Se matricularon un mayor número de estudiantes hombre (55,51 % vs 43,49 %) y 26 131 no presenta tener un

trabajo (98,29 %). Unos 11 663 estudiantes (43,87 %) ingresaron por el examen de Talento, dejando unos 6 585 (24,77 %), 2 407 (22,31 %) y 5 931 (9,05 %)de estudiantes que ingresaron por la Primera opción, por la academia pre-universitaria de la universidad y por programas especiales, respectivamente.

Con respecto al área de la primera matricula, 12 463 estudiantes (46, 88 %) eligieron Letras, 11 926 (44,86 %) Ciencias, 1 126 (3,33 %) Arquitectura y 292 (2 %) Educación. Asimismo, 21 989 (82, 72 %) de ellos provinieron de colegios particulares, 2 501 (9,40 %) de Nacionales, 1 289 (4,85 %) de parroquiales y 807 (3,03 %) de otros tipos de colegio. Sobre el origen de los colegios, 22 599 (85,01 %) pertenecen a Lima y Callao, 3 942 (14,82 %) a provincia y 45 (0,17 %) a fuera del país.

Con respecto a la información familiar, 25 758 (96,89 %) contaba con su madre viva al momento del ingreso y 22 960 (86,40 %) contaba con su padre vivo al momento del ingreso. Asimismo, 14 465 (54,41 %) de los estudiantes presentaban madres con grados academicos universitarios y 18 554 (69,78 %) presentaban padres con el mismo grado. La mediana del ingreso familiar es 4 323 $[RIC = 2750 - 6183]$ soles por familia al momento del ingreso (Tabla 5.2, 5.3 y 5.4).

Cuadro 5.2: Estadísticas descriptivas del enrolamiento (Parte I)

		Población N = 26 586
Factor	N	Porcentaje (%)
Evento		
Graduación	4 084	15,36
Abandono	3 764	14,16
Genero		
Hombre	15 024	55,51
Mujer	11 562	43,49
Indicador de Madre viva		
Si	25 758	96,89
No	3 617	13,60
Indicado de Padre vivo		
Si	22 969	86,40
No	3 617	13,60
Nivel de educación de la madre		
Universitaria	14 465	54,41
Técnica	5 357	20,15
Secundaria	6 231	23,44
Primaria	533	2

Cuadro 5.3: Estadísticas descriptivas del enrolamiento (Parte II)

	Población N = 26 586	
Factor	N	Porcentaje (%)
Nivel de educación del padre		
Universitaria	18 554	69,78
Técnica	3 400	12,79
Secundaria	4 330	16,29
Primaria	302	1,14
Trabaja		
Si	455	1,71
No	26 131	98,29
Tipo de admisión		
Ingreso por examen	11 663	43,87
Ingreso por examen - Primera opción	6 585	24,77
Ingreso por centro pre - universitario	2 407	9,05
Ingreso por programas especiales de admisión	5 931	22,31
Área de primera matricula		
Arquitectura	1 126	3,33
Letras	12 463	46,88
Educación	292	2
Ciencias	11 926	44,86
Arte	779	2,93
Tipo de escuela		
Particular	21 989	82,72
Nacional	2 501	9,40
Parroquial	1 289	4,85
Otros	807	3,03
Departamento del colegio de origen		
Lima y Callao	22 599	85,01
Provincia	3 942	14,82
Escuela extranjera	45	0,17
Ingreso familiar	4 323 [2 750 - 6 183]	

Cuadro 5.4: Estadísticas descriptivas del enrolamiento (Parte III)

		Población N = 26 586
Factor	N	Porcentaje (%)
Nuevos ingresos por semestre		
2004 - I	1 688	6,35
2004 - II	663	2,49
2005 - I	1 835	6,90
2005 - II	591	2,22
2006 - I	2 214	8,33
2006 - II	605	2,28
2007 - I	2 246	8,45
2007 - II	622	2,34
2008 - I	2 244	8,44
2008 - II	705	2,65
2009 - I	2 501	9,40
2009 - II	769	2,89
2010 - I	2 374	8,93
2010 - II	977	3,67
2011 - I	2 249	8,46
2011 - II	844	3,17
2012 - I	2 645	9,95
2012 - I	814	3,06

Una idea sobre el comportamiento general del evento respuesta de un estudiante en la muestra analizada puede ser observada mediante la estimación de la función de riesgo y supervivencia con el método de Kaplan - Meier. En la Tabla 5.5, se observa la estimación de las funciones antes mencionadas para el evento de abandono, en donde el 86,1 % de los estudiantes no presentaron el evento hasta en 3 años (6 semestres), el 81,6 % no presentaban el evento a los 5 años (10 semestres) fecha en que se deberían dar los eventos de graduación y un 50,9 % que no presentaba el evento después de 9 años (18 semestres).

En la Tabla 5.6, se observa el simil para el evento de graduación, a diferencia del evento de abandono los eventos comienzan a suceder a partir del 9no semestre, que es lo esperado en el tiempo de culminación de una carrera. Así el 66,8 % de los estudiantes no presentaban el evento a los 6 años (12vo semestre) y un 10 % de los alumnos no presento el evento después de 9 años (18 semestres).

El análisis confirma la separación cuasi - perfecta de los eventos a medida que el evento de graduación se presenta únicamente en los semestres posteriores a los 5 años de carrera, a diferencia del evento de abandono el cual se da en cualquier semestre del tiempo observado.

Cuadro 5.5: Estimador de Kaplan - Meier para el tiempo al abandono

Semestre	En riesgo	Fallas	Riesgo	Supervivencia	95 % IC
1er	26 586	5	0.00	1.00	(1.000, 1.000)
2do	24 914	724	0.029	0.971	(0.969 , 0.973)
3er	21 220	686	0.061	0.939	(0.936, 0.942)
4to	19 363	524	0.086	0.914	(0.910, 0.918)
5to	16 640	530	0.115	0.885	(0.881, 0.889)
6to	15 023	402	0.139	0.861	(0.856, 0.866)
7mo	12 685	311	0.160	0.840	(0.835, 0.845)
8vo	11 598	157	0.171	0.829	(0.823, 0.834)
9no	9 684	85	0.179	0.821	(0.816, 0.827)
10mo	8 913	62	0.184	0.816	(0.810, 0.821)
11vo	6 926	54	0.191	0.809	(0.803, 0.815)
12vo	5 465	58	0.199	0.801	(0.794, 0.807)
13vo	3 170	42	0.210	0.790	(0.783, 0.797)
14vo	1 911	54	0.232	0.768	(0.759, 0.777)
15vo	911	29	0.257	0.743	(0.731, 0.756)
16vo	433	26	0.301	0.699	(0.679, 0.719)
17vo	147	7	0.335	0.665	(0.635, 0.697)
18vo	34	8	0.491	0.509	(0.420, 0.617)

Cuadro 5.6: Estimador de Kaplan - Meier para el tiempo a la graduación

Semestre	En riesgo	Fallas	Riesgo	Supervivencia	95 % IC
9no	9 684	10	0.001	0.999	(0.998, 1.000)
10mo	8 913	380	0.044	0.956	(0.952, 0.961)
11vo	6 926	785	0.152	0.848	(0.840, 0.856)
12vo	5 465	1 160	0.332	0.668	(0.657, 0.679)
13vo	3 170	795	0.500	0.500	(0.488, 0.514)
14vo	1 911	535	0.640	0.360	(0.347, 0.374)
15vo	911	246	0.737	0.263	(0.249, 0.278)
16vo	433	128	0.815	0.185	(0.171, 0.201)
17vo	147	36	0.860	0.140	(0.124, 0.158)
18vo	34	9	0.887	0.103	(0.081, 0.130)

En base al análisis realizado, se estimó un modelo inicial con la metodología propuesta con el primer nivel de variables que estan relacionadas a las características de los estudiantes al momento de su ingreso. Como se muestra en la Tabla 5.7 y 5.8, la mayoría de variables utilizadas para la estimación muestran ser poco significativas a excepción de la condición del alumnos de haber ingresado por la primera opción. Por otro lado, se muestra desviaciones

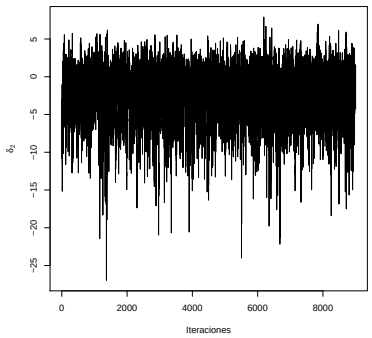
significativas en los intervalos de credibilidad asociados a los logodds basales posiblemente por iteraciones en las cuales los valores estimados alcanzaban valores extremadamente altos, dichos escenarios se observan en las cadenas de Markov y en las densidades de las distribuciones a posteriori para el log odds basal del periodo 2 y el coeficiente del género para ambos eventos (Figura 5.2).

Cuadro 5.7: Modelo de odds proporcionales Bayesiano para outcomes universitario (Parte I)

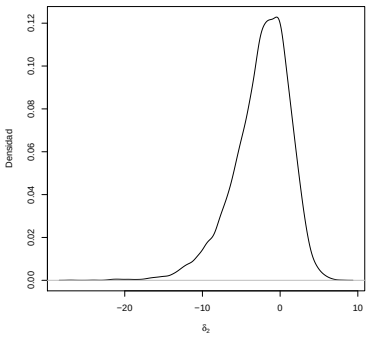
				Estimación Bayesiana
Factor	Evento 1 Graduacion	95 % IC	Evento 2 Abandono	95 % IC
Genero				
Mujer	-	-	-	-
Hombre	0.04	(-0.58, 0.58)	0.14	(-0.54, 0.88)
Indicador de Madre viva				
Si	-	-	-	-
No	-0.05	(-2.10, 1.36)	-0.04	(-1.57, 1.42)
Indicado de Padre vivo				
Si	-	-	-	-
No	-0.06	(-1.03,0.65)	-0.06	(-0.87,0.75)
Nivel de educación de la madre				
Universitaria	-	-	-	-
Tecnica	-0.05	(-0.83, 0.62)	0.04	(-0.78,1.17)
Secundaria	-0.35	(-1.55, 0.63)	0.09	(-0.80, 0.97)
Primaria	-0.16	(-5.75, 2.58)	0.07	(-2.73, 3.06)
Nivel de educación del padre				
Universitaria	-	-	-	-
Tecnica	-0.03	(-1.03, 0.69)	-0.09	(-0.88, 0.90)
Secundaria	-0.15	(-1.99, 0.72)	0.00	(-0.91, 0.90)
Primaria	-0.56	(-4.08, 1.89)	-0.45	(-5.17, 2.17)
Trabaja				
No	-	-	-	-
Si	-0.65	(-5.71, 1.36)	-0.34	(-2.84, 2.17)
Tipo de admisión				
Ingreso por examen	-	-	-	-
Ingreso por examen - Primera opción	0.85	(0.24, 1.39)	-1.25	(-2.41, -0.17)
Ingreso por centro pre - universitario	-0.16	(-1.93, 0.83)	0.02	(-0.96, 0.99)
Ingreso por programas especiales	-0.02	(-1.11, 1.02)	0.00	(-1.22, 0.84)
Ingreso por tercio superior	-0.07	(-1.19, 0.73)	-0.05	(-1.39, 0.82)

Cuadro 5.8: Modelo de odds proporcionales Bayesiano para outcomes universitario (Parte II)

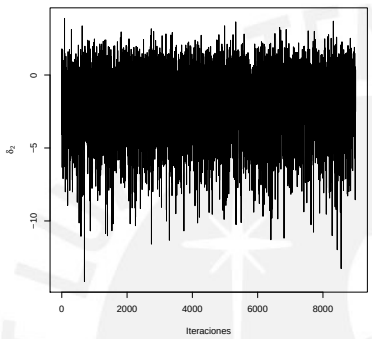
			Bayesian Estimation	
Factor	Evento 1 Graduacion	95 % IC	Evento 2 Abandono	95 % IC
Área de primera matricula				
Arquitectura	-	-	-	-
Letras	0.04	(-0.76, 0.75)	-0.34	(-1.17, 0.78)
Educación	-0.05	(-1.72, 1.51)	-0.22	(-2.65, 1.92)
Ciencias	0.12	(-0.47, 0.74)	0.51	(-0.55, 1.34)
Arte	-0.10	(-2.27, 1.11)	0.01	(-2.01, 1.37)
Tipo de escuela				
Particular	-	-	-	-
Nacional	-0.05	(-1.51, 0.69)	-0.03	(-1.16, 0.83)
Parroquial	0.00	(-1.14, 1.24)	-0.04	(-1.73, 0.92)
Otros	-0.03	(-1.34, 1.23)	0.01	(-1.51, 0.99)
Departamento del colegio de origen				
Lima y Callao	-	-	-	-
Provincia	0.05	(-0.63, 0.92)	0.00	(-1.05, 0.89)
Escuela extranjera	-3.90	(-22.95, 5.79)	-4.53	(-29.13, 4.26)
log odds				
2004 - I	-6.40	(-9.71, -4.09)	-4.21	(-5.77, -2.77)
2004 - II	-2.42	(-10.97, 3.25)	-2.01	(-7.01, 1.51)
2005 - I	-2.84	(-11.17, 2.65)	-0.13	(-2.85, 2.14)
2005 - II	-2.81	(-11.78, 2.63)	0.69	(-1.47, 2.64)
2006 - I	-2.97	(-10.86, 2.38)	0.81	(-1.02, 2.63)
2006 - II	-3.16	(-11.07, 2.35)	-0.55	(-3.25, 1.72)
2007 - I	-3.18	(-11.84, 1.92)	-2.83	(-7.15, 0.27)
2007 - II	-3.46	(-11.84, 1.92)	0.63	(-1.11, 2.35)
2008 - I	-3.41	(-11,42, 1.83)	0.73	(-0.88, 2.42)
2008 - II	1.55	(-1.30, 5.09)	-1.10	(-3.65, 1.04)
2009 - I	-3.13	(-10.91, 2.01)	0.20	(-1.50, 1.93)
2009 - II	2.69	(0.22, 5.96)	-0.09	(-1.87, 1.67)
2010 - I	3.10	(0.69, 6.43)	0.44	(-1.19, 2.10)
2010 - II	3.42	(1.03, 6.72)	-0.12	(-1.94, 1.66)
2011 - I	2.59	(0.09, 5.92)	0.86	(-0.67, 2.47)
2011 - II	2.66	(0.17, 5.99)	-0.62	(-2.73, 1.25)
2012 - I	2.77	(0.30, 6.12)	0.36	(-1.29, 2.06)
2012 - II	3.62	(1.28, 6.89)	0.88	(-0.64, 2.48)



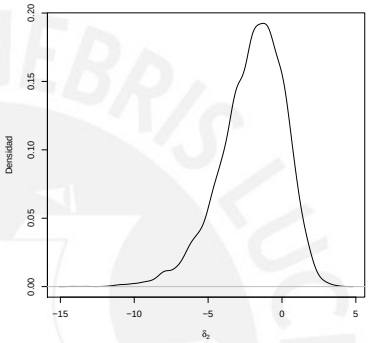
(a) logodds 2 del evento de graduación



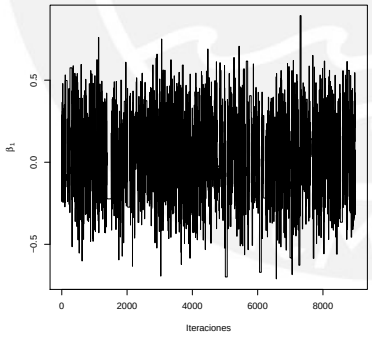
(b) logodds 2 del evento de graduación



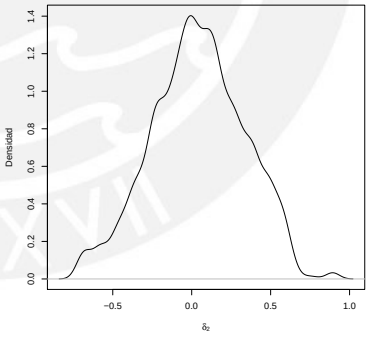
(c) logodds 2 del evento de abandono



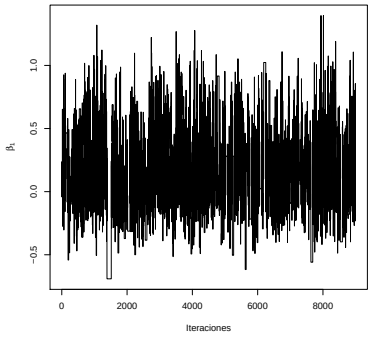
(d) logodds 2 del evento de abandono



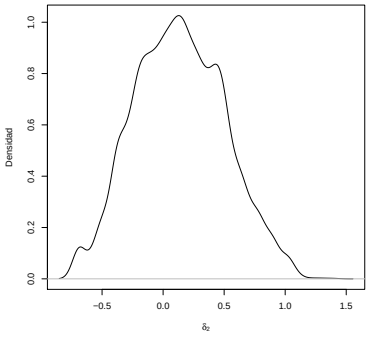
(e) Coeficiente asociado al género del evento de graduación



(f) Coeficiente asociado al género del evento de graduación



(g) Coeficiente asociado al género del evento de abandono



(h) Coeficiente asociado al género del evento de abandono

Figura 5.2: Cadenas de Markov y densidades de las distribuciones a posteriori del log odd basal del segundo periodo y el coeficiente de regresión asociado a la variable género.

Capítulo 6

Conclusiones

6.1. Conclusiones

En este trabajo se revisaron los modelos de riesgos competitivos bayesianos aplicado a una base de datos de la Pontificia Universidad Católica del Perú con el objetivo de predecir los eventos que un alumno se gradue o abandone la universidad. Asimismo, se mostró los problemas de estimación frecuentes en este tipo de data que resultan de la temporabilidad de ocurrencia de los eventos, donde la graduación de un alumno es frecuente a partir de 5 años de carrera de un alumno y el abandono en cambio resulta ser más frecuente al inicio de la carrera de este alumno.

Se utilizó como referencia el trabajo realizado por Vallejos y Steel (2017) en el cual se aplicó el mismo modelo para un conjunto de datos proveniente de la Pontificia Universidad Católica de Chile. Adicionalmente se detallaron los pasos metodológicos de la construcción del modelo bayesiano generando un apartado con demostraciones matemáticas que comprueban los supuestos detras del modelo.

Otro de los resultados importantes del trabajo fue el comparar los resultados de un modelo clásico con un modelo bayesiano de riesgos competitivos a fin de demostrar la mejora en la estimación y los problemas de sesgo ocasionados por la separación casi - perfecta de los eventos.

Los resultados de la aplicación a los datos mencionados resultaron en la no significancia de la mayoría de las variables incluidas en un modelo inicial con excepción de la variable que indica si el haber ingresado por la primera opción con respecto a haber ingresado por la vía tradicional la cual aumentaba la probabilidad de graduación y reducía la probabilidad de abandono. Asimismo, se observaron algunos sesgos en las estimaciones de los log odds basales resultantes de outliers en sus cadenas de Markov.

6.2. Sugerencias para investigaciones futuras

- Dentro del marco del modelo analizado se necesita incluir variables que cambien en el tiempo que estan presentes en el conjunto de datos provista por la universidad
- Mejorar la eficiencia del algoritmo reduciendo los tiempos de convergencia y los sesgos encontrados en las estimaciones realizadas.

Apéndice A

Resultados Complementarios

A.1. Representación de probabilidades binomiales bajo una distribución Polya - Gamma

Partimos del lado izquierdo de la identidad (3.11) el cual denota la probabilidad en un modelo logístico.

En primer lugar podemos desagregar el numerador de la forma:

$$\begin{aligned}\frac{(e^{\psi})^a}{(1+e^{\psi})^b} &= \frac{e^{\psi a}e^{-\frac{\psi b}{2}}e^{\frac{\psi b}{2}}}{(1+e^{\psi})^b} \\ &= e^{(a-\frac{b}{2})\psi}2^{-b}2^b(\frac{e^{\frac{\psi}{2}}}{1+e^{\psi}})^b \\ &= 2^{-b}e^{k\psi}(\frac{2e^{\frac{\psi}{2}}}{1+e^{\psi}})^b. \end{aligned} \tag{A.1}$$

Donde $k = a - \frac{b}{2}$. Asimismo, recordando la definición de coseno hiperbólico:

$$\begin{aligned}cosh(x) &= \frac{e^x + e^- x}{2} \\ &= \frac{e^{2x}+1}{2e^x} \\ &= \frac{1+e^{-2x}}{2e^{-x}}. \end{aligned} \tag{A.2}$$

Podemos escribir el término de la derecha de A.2 :

$$(\frac{2e^{\frac{\psi}{2}}}{1+e^{\psi}})^b = \frac{1}{cosh^b(\frac{\psi}{2})}. \tag{A.3}$$

Reemplazando:

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b}e^{k\psi}\frac{1}{cosh^b(\frac{\psi}{2})}. \tag{A.4}$$

Recordando la transformada de Laplace:

Sea una función de densidad $f(t)$ definida para todos los números postivos $t \geq 0$ entonces:

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t)dt = E[exp(-st)]. \tag{A.5}$$

La transformada de Laplace para la función de densidad que tiene como variable aleatoria ω que sigue una distribución polya - gamma de la forma $\omega \sim PG(b, 0), b > 0$:

$$E[exp(-\omega t)] = \prod_{k=1}^{\infty}(1 + \frac{t}{2\pi^2(k - \frac{1}{2})^2})^- b = \frac{1}{cosh^b(\sqrt{\frac{t}{2}})}. \tag{A.6}$$

Reemplazando en la función

$$\frac{(e^\psi)^a}{(1 + e^\psi)^b} = 2^{-b}e^{k\psi}E[exp(-\omega\frac{\psi^2}{2})] = 2^{-b}e^{k\psi}\int_0^\infty e^{-\omega\frac{\psi^2}{2}}p(\omega)d\omega. \tag{A.7}$$

Donde $\omega \sim PG(b, 0), b > 0$

Si tratamos esta expresión como la distribución conjunta de (ψ, ω) podemos derivar la condicional:

$$p(\omega|\psi) = \frac{e^{-\omega\frac{\psi^2}{2}}p(\omega)}{\int_0^\infty e^{-\omega\frac{\psi^2}{2}}p(\omega)d\omega}. \tag{A.8}$$

Donde $p(\omega)$ es la densidad de $PG(b, 0)$.

Específicamente la función de probabilidad de una variable aleatoria $PG(b, c)$ sigue:

$$p(x|b, c) = \frac{exp(-\frac{c^2}{2}x)p(x|b, 0)}{E\left\{exp(-\frac{c^2}{2}x)\right\}}. \tag{A.9}$$

Comparando con la probabilidad condicional encontrada, si reemplazamos $\psi = c$ podemos ver que $\omega|\psi \sim PG(b, \psi)$

Apéndice B

Código en R

```
##########################################
# I. Simulaciones
##########################################

# I.1 Escenario A:
####################

# I.1.1 Simulaci\'on de datos
#################################

## Funciones auxiliares

##  Construcci\'on de la matriz de dise\~no
X.design<-function(X.Period,inc,cvar=1,nvar=3)
{
  X=X.Period
  for(i in 1:cvar){
    if(inc[i]==1){X=cbind(X,get(paste0("x",i,1)),get(paste0("x",i,2)))}
  }
  a = cvar+1
  for(i in a:nvar){
    if(inc[i]==1){X=cbind(X,get(paste0("x",i)))}
  }
  return(X)
}
## Indicador de que efectos estan activos
ind.var<-function(inc,n.period=16,cvar=1,nvar=3)
{
  aux =1:n.period
  cont=n.period

  for(i in 1:cvar){
```

```
    if(inc[i]==1){aux=c(aux,cont+i,cont+i+1)}
    cont = cont+1
  }
  a = cvar+1
  for(i in a:nvar){
    if(inc[i]==1){aux=c(aux,cont+i)}
  }
  return(aux)
}


# n           : # Students
# nvar        : # Vars
# cvar        : # categorical variables
# Exc         : relevant variables indicator
DatSim   = function(n = 400,nvar = 3,cvar = 1,Exc = c(1,1,0),
n.period = 16,n.out=3){

  ## Simulando covariables para n estudiantes
  n.students=n

  ### Variables categoricas (C = 3)

  if(cvar!= 0){

    for(i in 1:cvar){

      assign(paste0("x",i),rbinom(n=n.students,size=2,prob=0.5))

    }
  }

  ### Variables dicotomicas (C = 2)

  dvar = nvar - cvar
  a = cvar+1
  for(j in a:nvar){

    assign(paste0("x",j),rbinom(n=n.students,size=1,prob=0.5))

  }
```

```
### Variables relevantes

X.students = data.frame(1:n)

for(i in 1:nvar){

  X.students = data.frame(X.students,factor(get(paste0("x",i))))

}
X.students          = X.students[-1]
names(X.students)   = paste0("x",1:nvar)

if(is.null(Exc)== FALSE){

  X.students_aux = X.students[,Exc]

}else{X.students_aux = X.students}

### Indicadores de la simulaci\'on

n.var=dim(X.students_aux)[2]
if(is.null(n.var)==TRUE){n.var=1}

#Dfaux       = data.frame(X.students)

X.students = model.matrix(~.,X.students)[,-1]
n.effects  = dim(X.students)[2]
if(is.null(n.effects)==TRUE){n.effects=1}

## Numero de posibles eventos (excluyendo la censura)
n.outcomes=n.out

## Indicador del n\'umero de periodos

n.deltas=n.period

## Valor de los par\'ametros ( Primer delta representa un intercepto)

delta=matrix(0,ncol=n.out,nrow=n.deltas)
delta[,1]<-c(-8.23,-5.62,-4.82,-4.88,-5.96,-5.76,
                    -4.89, 4.37, 4.06, 5.98, 5.95, 6.58,
                      6.93, 7.34, 5.47, 6.40)
```

```
delta[,2]<-c(-8.51, 6.07, 5.42, 4.89, 4.81, 4.30,
                    4.46, 4.96, 4.37, 3.54,-4.52,-4.92,
                   -3.90,-4.35,-3.70,-4.48)
delta[,3]<-c(-3.14, 1.31, 0.28, 0.77,-0.80, 0.45,
                   -2.22, 0.64, 0.24,-0.86,-0.06, 0.76,
                    0.34,-0.14, 1.85,-0.57)

beta=matrix(0,ncol=n.out,nrow=n.effects)
beta[,1]<-c(0,1,-1,0)
beta[,2]<-c(1,0,0,0)
beta[,3]<-c(0.5,-0.5,0,0)

## Cociente de riesgo de Causa - Especifica en casa tiempo

hr1<-matrix(0,ncol=n.deltas,nrow=n.students)
hr2<-matrix(0,ncol=n.deltas,nrow=n.students)
hr3<-matrix(0,ncol=n.deltas,nrow=n.students)

aux1<-delta[,1]+c(0,rep(delta[1,1],times=n.deltas-1))
aux2<-delta[,2]+c(0,rep(delta[1,2],times=n.deltas-1))
aux3<-delta[,3]+c(0,rep(delta[1,3],times=n.deltas-1))

for(i in 1:n.students)
{
  hr1[i,]<-exp(aux1+rep(as.numeric(X.students[i,]%*%beta[,1]),n.deltas))
  hr2[i,]<-exp(aux2+rep(as.numeric(X.students[i,]%*%beta[,2]),n.deltas))
  hr3[i,]<-exp(aux3+rep(as.numeric(X.students[i,]%*%beta[,3]),n.deltas))
}

h0<-1/(1+hr1+hr2+hr3)
h1<-hr1/(1+hr1+hr2+hr3)
h2<-hr2/(1+hr1+hr2+hr3)
h3<-hr3/(1+hr1+hr2+hr3)

## Matriz auxiliar para la creaci\'on de variables dummy para indicadores de

delta.aux<-matrix(0,ncol=n.deltas,nrow=n.deltas)
delta.aux[,1]<-rep(1,times=n.deltas)
for(i in 2:n.deltas){delta.aux[i,i]=1}

## Simulacion de los eventos para cada periodo/estudiante
```

```
  data.long<-NULL; period=0
  for(i in 1:n.students)
  {
    outcome=0; period=0
    while(outcome==0 & period<n.deltas)
    {
      period=period+1
      outcome<-sample(x=c(0,1,2,3),size=1,replace=TRUE,
      prob=c(h0[i,period],h1[i,period],h2[i,period],
      h3[i,period]))
      data.long<-rbind(data.long,c(i,outcome,delta.aux[period,],
                                   X.students[i,]))
    }
  }

  return(data.long)
}

for(i in 1:M){
  A=DatSim(400,3,1,Exc = c(1,1,0),n.period=16,n.out=3)
  Y=A[,2]
  X=as.matrix(A[,-c(1,2)])

# I.1.2 Estimador de Kaplan - Meier
#########################################

  ### Cociente de riesgo para cada evento

  #### Primer evento

  B1= cbind(A,ave(A[,2], A[,1], FUN = seq_along))
  B1= cbind(B1,as.numeric(I(A[,2]==1)))

  Surv(B1[,23],B1[,24])
  km  = survfit(Surv(B1[,23],B1[,24])~1)
  res = summary(km)

  #### Segundo evento

  B2= cbind(A,ave(A[,2], A[,1], FUN = seq_along))
  B2= cbind(B2,as.numeric(I(A[,2]==2)))
```

```
  Surv(B2[,23],B2[,24])
  km  = survfit(Surv(B2[,23],B2[,24])~1)
  res = summary(km)

  #### Tercer evento

  B3= cbind(A,ave(A[,2], A[,1], FUN = seq_along))
  B3= cbind(B3,as.numeric(I(A[,2]==3)))

  Surv(B3[,23],B3[,24])
  km  = survfit(Surv(B3[,23],B3[,24])~1)
  res = summary(km)

# I.1.3 Estimaci\'on por el m\'etodo Frecuentista
#####################################################

  model1 = multinom ( Y ~X-1 , Hess = TRUE)
  summary(model1)

# I.1.4 Estimaci\'on por el m\'etodo Bayesiano
##################################################

  ### Parametros requeridos

  k=dim(X)[2]
  n.outcomes=3
  beta0=array(0,dim=c(1,k,n.outcomes))
  mean.beta=array(0,dim=c(1,k,n.outcomes))

  ### MCMC Cadena

  n.deltas=16
  chain=MCMC.MLOG(N=100000,thin=10,Y=Y,X=X,t0=n.deltas,beta0=beta0,
  mean.beta=mean.beta,prec.delta=1/100,df.delta=1,
  logg0=c(5,5,5),ls.g0=c(3,3,3),prior="Benchmark-Beta",
  ar=0.44,fix.g=FALSE, ncov = 3)


# I.2 Simulaci\'on de datos Escenario B:
################################################

DatSim_mu   = function(n = 400,nvar = 3,cvar = 1,Exc = c(1,1,0),
```

```
                                        n.period = 16,n.out=3){

 n.students=n

if(cvar!= 0){

  for(i in 1:cvar){

    assign(paste0("x",i),rbinom(n=n.students,size=2,prob=0.5))

  }
}

dvar = nvar - cvar
a = cvar+1
for(j in a:nvar){

  assign(paste0("x",j),rbinom(n=n.students,size=1,prob=0.5))

}


X.students = data.frame(1:n)

for(i in 1:nvar){

  X.students = data.frame(X.students,factor(get(paste0("x",i))))

}
X.students          = X.students[-1]
names(X.students)   = paste0("x",1:nvar)

if(is.null(Exc)== FALSE){

  X.students_aux = X.students[,Exc]

}else{X.students_aux = X.students}

n.var=dim(X.students_aux)[2]
if(is.null(n.var)==TRUE){n.var=1}

#Dfaux        = data.frame(X.students)
```

```
  X.students = model.matrix(~.,X.students)[,-1]
  n.effects  = dim(X.students)[2]
  if(is.null(n.effects)==TRUE){n.effects=1}

  n.outcomes=n.out

  n.deltas=n.period

   delta=matrix(0,ncol=n.out,nrow=n.deltas)

# Graduacion
  delta[,1]<-c(-6.23,-5.62,-4.82,-4.88,-5.96,-5.76,-4.89, 4.37,
                        4.06, 5.98, 5.95, 6.58, 6.93, 7.34, 5.47, 6.40)

# Abandono involuntario
  delta[,2]<-c(-9.51, 6.07, 5.42, 4.89, 4.81, 4.30, 4.46, 4.96,
                       4.37, 3.54,-4.52,-4.92,-3.90,-4.35,-3.70,-4.48)

# Abandono voluntario
  delta[,3]<-c(-5.14, 1.31, 0.28, 0.77,-0.80, 0.45,-2.22, 0.64,
                        0.24,-0.86,-0.06, 0.76, 0.34,-0.14, 1.85,-0.57)

  beta=matrix(0,ncol=n.out,nrow=n.effects)
  beta[,1]<-c(0,1,-1,0)
  beta[,2]<-c(1,0,0,0)
  beta[,3]<-c(0.5,-0.5,0,0)



  hr1<-matrix(0,ncol=n.deltas,nrow=n.students)
  hr2<-matrix(0,ncol=n.deltas,nrow=n.students)
  hr3<-matrix(0,ncol=n.deltas,nrow=n.students)

  aux1<-delta[,1]+c(0,rep(delta[1,1],times=n.deltas-1))
  aux2<-delta[,2]+c(0,rep(delta[1,2],times=n.deltas-1))
  aux3<-delta[,3]+c(0,rep(delta[1,3],times=n.deltas-1))

  for(i in 1:n.students)
  {
    hr1[i,]<-exp(aux1+rep(as.numeric(X.students[i,]%*%beta[,1]),n.deltas))
    hr2[i,]<-exp(aux2+rep(as.numeric(X.students[i,]%*%beta[,2]),n.deltas))
    hr3[i,]<-exp(aux3+rep(as.numeric(X.students[i,]%*%beta[,3]),n.deltas))
```

```
  }


  h0<-1/(1+hr1+hr2+hr3)
  h1<-hr1/(1+hr1+hr2+hr3)
  h2<-hr2/(1+hr1+hr2+hr3)
  h3<-hr3/(1+hr1+hr2+hr3)


  delta.aux<-matrix(0,ncol=n.deltas,nrow=n.deltas)
  delta.aux[,1]<-rep(1,times=n.deltas)
  for(i in 2:n.deltas){delta.aux[i,i]=1}


  data.long<-NULL; period=0
  for(i in 1:n.students)
  {
    outcome=0; period=0
    while(outcome==0 & period<n.deltas)
    {
      period=period+1
      outcome<-sample(x=c(0,1,2,3),size=1,replace=TRUE,
      prob=c(h0[i,period],h1[i,period],h2[i,period],
      h3[i,period]))
      data.long<-rbind(data.long,c(i,outcome,delta.aux[period,],
                                   X.students[i,]))
    }
  }

  return(data.long)
}


  for(i in 1:M){
    A=DatSim(400,3,1,Exc = c(1,1,0),n.period=16,n.out=3)
    Y=A[,2]
    X=as.matrix(A[,-c(1,2)])

    # I.2.2 Estimador de Kaplan - Meier
    ########################################

    ### Cociente de riesgo para cada evento
```

```
#### Primer evento

B1= cbind(A,ave(A[,2], A[,1], FUN = seq_along))
B1= cbind(B1,as.numeric(I(A[,2]==1)))

Surv(B1[,23],B1[,24])
km  = survfit(Surv(B1[,23],B1[,24])~1)
res = summary(km)

#### Segundo evento

B2= cbind(A,ave(A[,2], A[,1], FUN = seq_along))
B2= cbind(B2,as.numeric(I(A[,2]==2)))

Surv(B2[,23],B2[,24])
km  = survfit(Surv(B2[,23],B2[,24])~1)
res = summary(km)

#### Tercer evento

B3= cbind(A,ave(A[,2], A[,1], FUN = seq_along))
B3= cbind(B3,as.numeric(I(A[,2]==3)))

Surv(B3[,23],B3[,24])
km  = survfit(Surv(B3[,23],B3[,24])~1)
res = summary(km)

# I.2.3 Estimaci\'on por el m\'etodo Frecuentista
################################################

model1 = multinom ( Y ~X-1 , Hess = TRUE)
summary(model1)

# I.2.4 Estimaci\'on por el m\'etodo Bayesiano
##############################################

### Parametros requeridos

k=dim(X)[2]
n.outcomes=3
beta0=array(0,dim=c(1,k,n.outcomes))
mean.beta=array(0,dim=c(1,k,n.outcomes))
```

```
### Cadena de markov

n.deltas=16
chain=MCMC.MLOG(N=100000,thin=10,Y=Y,X=X,t0=n.deltas,beta0=beta0,
mean.beta=mean.beta,prec.delta=1/100,df.delta=1,
logg0=c(5,5,5),ls.g0=c(3,3,3),prior="Benchmark-Beta",
ar=0.44,fix.g=FALSE, ncov = 3)


##########################################
# II. Aplicaci\'on
##########################################

# Modelo Bayesiano
## Parametros requeridos
set.seed(1234)
Y = dt.fm[,2]
X = as.matrix(dt.fm[,-c(1,2)])
k=dim(X)[2]
n.outcomes=2
beta0=array(0,dim=c(1,k,n.outcomes))
mean.beta=array(0,dim=c(1,k,n.outcomes))

## Cadena de Markov

n.deltas=18
chain=MCMC.MLOG(N=100000,thin=10,Y=Y,X=X,t0=n.deltas,beta0=beta0,
mean.beta=mean.beta,prec.delta=1/100,df.delta=1,
logg0=c(5,5),ls.g0=c(3,3),prior="Benchmark-Beta",
ar=0.44,fix.g=FALSE, ncov = 24,cvar=0,nvar=24)
}
```

Bibliografía

Agresti, A. (2003). *Categorical data analysis*, Vol. 482, John Wiley & Sons.

Albert, A. y Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models., *Biometrika* (71): 1–10.

Albert, J. y Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**: 669–679.

Allison, P. (1982). Discrete-time methods for the analysis of event histories, pp. 61–98.

Bedrick, E., Christensen, R. y Johnson, W. (1996). A new perspective on priors for generalized linear models, *J. Amer. Statist. Assoc.* **91**: 1450–1460.

Bian, G. y Dickey, J. (1991). Properties of multivariate cauchy and poly-cauchy distributions with bayesian g-prior applications, *Technical Report No. 567* (567).

Clarke, P., Gray, A., Briggs, A., Farmer, A., Fenn, P., Stevens, R., Matthews, D., Stratton, I., Holman, R., Group, U. P. D. S. U. et al. (2004). A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the united kingdom prospective diabetes study (ukpds) outcomes model (ukpds no. 68), *Diabetologia* **47**(10): 1747–1759.

Cox, D. (1972). Regresion models and life - tables (with discussion), *J. R. Statist. Soc.* **B,34**: 187–220.

Crowder, M. (1996). On assessing independence of competing risks when failure times are discrete, *Liftim. Data Anal* **2**: 195–209.

Cui, W. y George, E. (2008). Emperical bayes vs. fully bayes variable selection, *Statistical Planning and Inference* **138**: 388–396.

DesJardins, S. L., Ahlburg, D. A. y McCall, B. P. (2002). A temporal investigation of factors related to timely degree completion, *The Journal of Higher Education* **73**(5): 555–581.

DesJardins, S. L., Ahlburg, D. A. y McCall, B. P. (2006). The effects of interrupted enrollment on graduation from college: Racial, income, and ability differences, *Economics of Education Review* **25**(6): 575–590.

Früwirth-Schnatter, S., FrÃ¼wirth, R., Held, L. y Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical model of non-gaussian data, *Statistics and Computing* **19**: 479–492.

Gelman, A., Jakulin, A., Pittau, M. y Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models, *Ann. Appl. Statist.* **2**: 1360–1383.

Gramacy, R. y Polson, N. (2012). Simulation-based regularized logistic regression, *Bayesian Analysis* **7**: 567–590.

Holmes, C. y Held, L. (2006). Bayesian auxilary variable models for binary and multinomial regression, *Bayesian Analysis* **1**: 145–168.

Hosmer, D. W., Lemeshow, S. y May, S. (2011). *Applied survival analysis*, Wiley Blackwell.

Lassibille, G. y Navarro Gómez, L. (2008). Why do higher education students drop out? evidence from spain, *Education Economics* **16**(1): 89–105.

Liang, F., Paulo, R., Molina, G., Clyde, M. y Berger, J. (2008). Mixtures of g priors for bayesian variable selection, *J. Am. Statist. Ass* **103**: 410–423.

Murtaugh, P., Burns, L. y Schuste, J. (1999). Predicting the retention of university students, *Res. Highr Educ.* **40**: 355–371.

Neethling, L. (2015). The determinants of academic outcomes: A competing risks approach, *School of Economics of University of Cape Town* .

Polson, N., Scott, J. y Windle, J. (2013). Bayesian inference for logistic models using polya-gamma latent variables, *J. Am. Statist. Ass.* **108**: 1339–1349.

Roberts, G. O. y Rosenthal, J. S. (2009). Examples of adaptive mcmc, *Journal of Computational and Graphical Statistics* **18**(2): 349–367.

Sabanés, B. y Held, L. (2011). Hyper-g priors for generalized linear models, *Baysn Anal.* **6**: 387–410.

Scott, M. y Kennedy, B. (2005). Pitfalls in pathways: some perspectives on competing risks event history analysis in education research, *J. Educ. Behav. Statist.* **30**: 413–442.

Singer, J. y Willet, J. (1993). It's about time: using discrete-time survival analysis to study duration and the timing of events, *J. Educ. Behav. Statist* **18**: 155–195.

Takam, R., Bezak, E. y Yeoh, E. (2009). Risk of second primary cancer following prostate cancer radiotherapy: Dvh analysis using the competitive risk model, *Physics in Medicine & Biology* **54**(3): 611.

Trujillo, P. y Raúl, M. (2016). An application of discrete time survival models to analyze student dropouts at a private university in peru.

Vallejos, C. y Steel, M. (2017). Bayesian survival modelling of university outcomes, *Journal of the Royal Statistical Society* **180**: 613–631.

Witte, J., Greenland, S. y Kim, L. (1998). Software for hierarchical modeling of epidemiologic data, *Epidemiology* **9**: 563–566.

Wolbers, M., Koller, M. T., Witteman, J. C. y Steyerberg, E. W. (2009). Prognostic models with competing risks: methods and application to coronary risk prediction, *Epidemiology* **20**(4): 555–561.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions, pp. 233–243.