

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO



SISTEMA DE TARIFACIÓN BONUS-MALUS
PARA LA RAMA DE SEGUROS DE AUTOMÓVIL

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN
ESTADÍSTICA

Presentado por:

Vivanco Ortiz, Yoshi Abel

Asesor: Dr. Valdivieso Serrano, Luis Hilmar

Miembros del jurado:

Dr. Bayes Rodriguez, Cristian Luis

Dr. Valdivieso Serrano, Luis Hilmar

Dr. Quiroz Cornejo, Zaida Jesus

Lima, Noviembre 2020

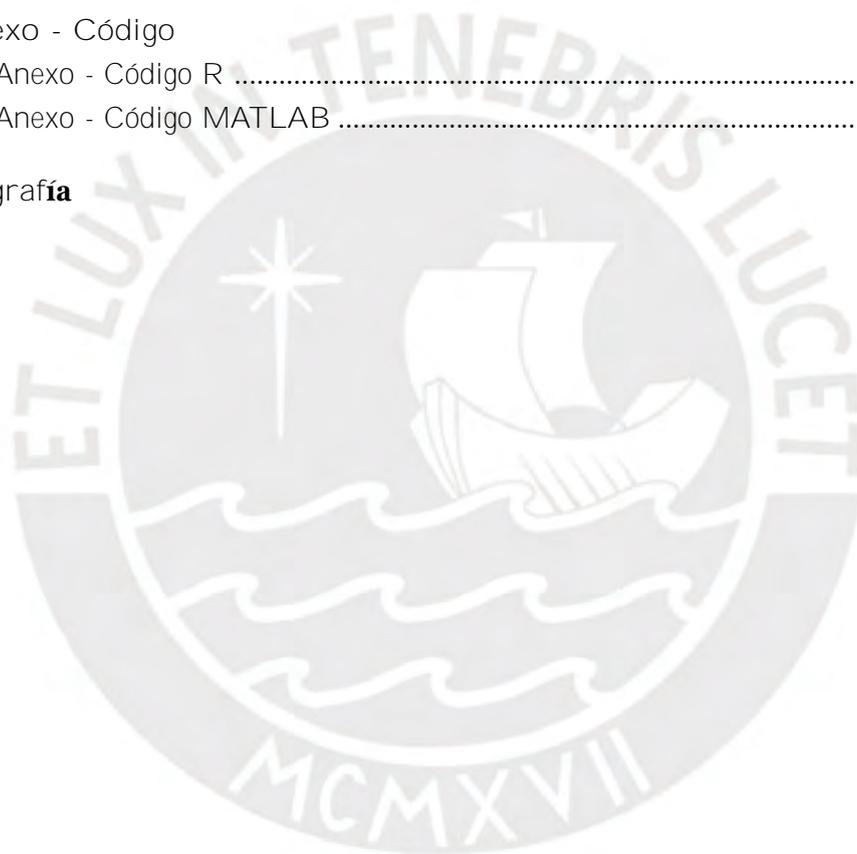
RESUMEN

En la actualidad, las empresas aseguradoras cuentan con productos de seguros cada vez más personalizados a las características de sus asegurados, de modo que, cada asegurado no pague el mismo monto de prima sino un monto proporcional a su comportamiento y perfil de riesgo. Una de las formas de atender esta necesidad de personalización en la tarificación es el Sistema Bonus-Malus (SBM), el cual ajusta una prima base considerando la historia de siniestros reportados por cada asegurado. En ese sentido, una historia sin siniestros crea bonificaciones (bonus) y por ende una reducción en la prima de seguro; y, una historia con siniestros genera penalizaciones (malus) y por ende un incremento en la prima de seguro. Por tanto, el objetivo de esta tesis es aplicar los modelos SBM basados en la frecuencia para un seguro de tipo vehicular. Para ello, en base a la información disponible de los asegurados, se construye un modelo de frecuencia de siniestros usando un GLM (Poisson, Binomial Negativa y sus variantes inflacionadas en ceros), cada modelo permite obtener una prima base y clases de riesgo basados en características heterogéneas. Luego, se comparan todos los modelos obtenidos para seleccionar el mejor ajuste para los datos analizados. Por último, se aplica el SBM y se determina en qué nivel se clasifica a cada asegurado en función al número de siniestros que reporte en el periodo de análisis, de esa manera, se determina el valor de la prima ajustada para cada asegurado. En resumen, este trabajo desarrolla un SBM con información a priori y a posteriori que permite obtener primas más justas para los asegurados de un producto de seguros vehiculares, de modo que, el asegurado que presente un comportamiento sin siniestros reportados pagará menos que un asegurado que presente siniestros en el periodo evaluado.

Índice general

| | |
|---|----|
| Índice de figuras | v |
| Índice de cuadros | vi |
| 1. Introducción | 1 |
| 1.1. Contexto | 1 |
| 1.2. Objetivos | 2 |
| 1.3. Metodología | 3 |
| 1.4. Organización del trabajo | 3 |
| 2. Conceptos previos | 5 |
| 2.1. Modelos lineales generalizados (GLM) | 7 |
| 2.1.1. Regresión de Poisson | 8 |
| 2.1.2. Regresión binomial negativa | 9 |
| 2.1.3. Regresión de Poisson inflacionada en ceros | 11 |
| 2.1.4. Regresión binomial negativa inflacionada en ceros | 11 |
| 2.2. Herramientas de diagnóstico y bondad de ajuste | 12 |
| 2.2.1. Devianza | 12 |
| 2.2.2. Test de Equidispersión | 13 |
| 2.2.3. Test de Vuong | 13 |
| 2.2.4. AIC | 14 |
| 2.2.5. BIC | 14 |
| 2.3. Cadenas de Markov | 15 |
| 3. Sistema Bonus-Malus | 18 |
| 3.1. Clasificación a priori | 19 |
| 3.2. Clasificación a posteriori | 19 |
| 3.2.1. Componentes principales de un SBM | 20 |
| 3.2.2. Modelo de credibilidad del SBM propuesto | 21 |
| 3.3. Escala del Sistema Bonus-Malus propuesto | 22 |
| 3.4. Reglas de transición, probabilidades de transición y distribución estacionaria | 23 |
| 3.5. Relatividades | 28 |

| | |
|--|----|
| 4. Aplicación | 32 |
| 4.1. Descripción del conjunto de datos utilizado | 32 |
| 4.2. Clasificación a priori | 33 |
| 4.2.1. Selección del modelo de regresión | 34 |
| 4.2.2. Selección de covariables más significativas | 35 |
| 4.2.3. Determinación de las clases de riesgo | 37 |
| 4.2.4. Determinación de prima a priori | 37 |
| 4.3. Clasificación a posteriori | 39 |
| 5. Conclusiones | 42 |
| 5.1. Conclusiones | 42 |
| 5.2. Recomendaciones | 43 |
| A. Anexo - Código | 44 |
| Anexo - Código R | 44 |
| Anexo - Código MATLAB | 47 |
| Bibliografía | 49 |



Índice de figuras

| | |
|------------------------|----|
| 3.1. Niveles SBM | 23 |
|------------------------|----|



Índice de cuadros

| | |
|--|----|
| 3.1. Niveles SBM - Vista Funcional | 25 |
| 4.1. Diccionario de las variables..... | 32 |
| 4.2. Volumen de siniestros reportados | 33 |
| 4.3. Análisis univariado del conjunto de datos | 33 |
| 4.4. Criterio de selección por mejor ajuste | 34 |
| 4.5. Resultados AIC y BIC para los modelos GLM en comparación | 34 |
| 4.6. Resultado Test de Vuong..... | 35 |
| 4.7. Significancia de covariables del modelo inicial..... | 36 |
| 4.8. Significancia de covariables del modelo reducido | 36 |
| 4.9. Clase de riesgo - Escenario 1..... | 38 |
| 4.10. Clase de riesgo - Escenario 2..... | 38 |
| 4.11. Resumen de resultados de las relatividades obtenidas por el SBM propuesto | 40 |
| 4.12. Aplicación práctica con el modelo propuesto | 40 |

Capítulo 1

Introducción

1.1. Contexto

En la actualidad nuestro país es visto por tener una pobre cultura de seguros y prevención, lo que nos lleva a tener una gran oportunidad en el desarrollo del sector asegurador en todos los niveles socio-económicos (Andrea Camargo, 2014). Esta oportunidad deberá estar asociada a productos de seguros personalizados que permitan al asegurado percibir un valor agregado (tarifa diferenciada, protección máxima y ahorro visible) y así pagar una prima que esté asociada a sus propias características o al histórico de sus siniestros, de modo que el asegurado no pague el mismo monto de prima que otros asegurados cuyas características o riesgos son diferentes. Esto se aprecia actualmente en el mercado de seguros de automóvil peruano, por ejemplo, durante el año 2017 y 2018 reconocidas empresas aseguradoras en el Perú lanzaron al mercado sus primeros productos de seguro vehicular (con y sin GPS), que incluyen el estudio del comportamiento del cliente o asegurado y cuya idea principal se basa en que no todas las personas tienen que pagar lo mismo, dado que no todas conducen de igual forma; de modo que, se premia a los que mejor manejen y se penaliza en caso contrario (Antonio y Van Berkum, 2015).

El ejemplo expuesto en el párrafo anterior muestra una forma de abordar la oportunidad existente en el sector asegurador de nuestro país, la cual consiste en el desarrollo de un tipo específico de sistema de tarifación basado en la experiencia denominado “Escala Bonus-Malus” o “Sistema de Tarifación Bonus-Malus”, en adelante SBM. Este tipo de sistema es ampliamente usado en la rama de seguros vehiculares, sobre todo en Europa y Norteamérica.

El SBM a estudiar es caracterizado por un monto de prima base que será ajustado en base a la historia de siniestros reportados por el asegurado. Una historia sin siniestros creará “bonificaciones (bonus)” y por ende una reducción en su prima de seguro; en cambio, una historia con siniestros generará “penalizaciones (malus)” y por ende un incremento en su prima de seguro. Cabe resaltar que los SBM son atractivos desde un punto de vista comercial y de transparencia al usuario final, debido a que facilita la explicación por parte de la compañía de seguros al asegurado sobre cómo sus siniestros reportados en un año determinado permitirán cambiar su prima de seguro inicial (prima base) en el año siguiente (prima ajustada).

Los SBM han evolucionado en el tiempo, estos se usan desde hace más de 70 años. Por ello, es preciso resaltar la literatura que ha marcado un hito en esta evolución, desde los SBM propuestos por [Lemaire \(1995\)](#) y [Denuit M. \(2007\)](#) basados en la frecuencia de los siniestros; pasando por el SBM propuesto por [Pinquet \(1997\)](#) basado en la severidad; hasta los SBM propuestos por [Pinto \(2014\)](#), [Tzougas y Frangos \(2007\)](#) y [Gevorg \(2016\)](#) que consideran información a priori y a posteriori. En ese sentido, entiéndase información a priori como la información con la que la empresa aseguradora cuenta antes de que el asegurado compre y use el seguro, por ejemplo, en el sector de seguros vehiculares se tiene: número de siniestros de los otros asegurados de la empresa aseguradora, datos personales del nuevo asegurado (edad, sexo, dirección), o características de su vehículo (tipo, valor monetario, potencia de motor), o sus datos socio-demográficos (densidad poblacional, número de accidentes en la región o provincia, y localidad o distrito donde vive el asegurado); la cual permite dividir la cartera de clientes (asegurados) en grupos o clases homogéneas, es decir, características similares o perfil de riesgo común. Asimismo, entiéndase información a posteriori como la historia de siniestros del asegurado, es decir, el número y/o monto de siniestros reportados (frecuencia y/o severidad) del nuevo asegurado durante el tiempo que se encuentra haciendo uso del seguro adquirido en la empresa aseguradora.

Por otra parte, si bien las variables a priori (relacionadas al concepto de información a priori indicado en el párrafo anterior) permiten dividir la cartera heterogénea en grupos o clases de riesgo homogéneas de asegurados que paguen una misma prima por cada clase diferenciada, no es posible observar algunos otros factores a priori relevantes, tales como los reflejos del conductor, su estilo de conducción, su consumo del alcohol durante el manejo, su experiencia de manejo, entre otros. Es decir, los grupos que se logran obtener con ayuda de las variables a priori observables aún presentan heterogeneidad residual. En ese sentido, la información a posteriori (historial de reclamaciones) permite revelar tales variables a priori no observables a través del número de siniestros reportados por los asegurados en el periodo analizado.

El SBM propuesto en este trabajo considera tanto información a priori como información a posteriori, es decir, inicialmente se clasificará al asegurado en una clase de riesgo, cuyo monto de prima base es diferenciada respecto de las otras clases de riesgo de la empresa aseguradora, y luego esta prima base se ajustará considerando el número de siniestros reportados por el asegurado en el periodo analizado, es decir, esta prima ajustada se determinará en función a la prima base (a priori) y la siniestralidad del asegurado en el periodo de estudio (a posteriori).

1.2. Objetivos

El objetivo general de esta tesis es estudiar y aplicar los modelos de tarificación Bonus-Malus basados en la frecuencia, que incorpore tanto información a priori como a posteriori de la experiencia de siniestros de los usuarios en un seguro de tipo

vehicular. De manera **específica**, se busca:

- Revisar la literatura acerca de los sistemas bonus-malus en seguros del tipo vehicular.
- Estudiar los modelos lineales generalizados (GLM) para aplicar los mismos a un sistema de tarificación.
- Estudiar las escalas bonus-malus y sus transiciones entre niveles **v**acadenas de Markov.
- Plantear y estudiar el sistema de tarificación bonus-malus con los GLM para la modelación del número de siniestros.
- Aplicar los modelos estudiados a datos reales para los fines de la presente tesis.

1.3. Metodología

A continuación se resume el enfoque metodológico de la presente tesis con el objetivo de entender cómo se relacionan cada uno de los conceptos asociados al modelo propuesto:

En base a la información disponible de la cartera de asegurados, tales como: género, edad, residencia, tipo de **vehículo**, valor de **vehículo**, o incluso el número de siniestros reportados por cada uno de los clientes que conforma la cartera; se construirá un modelo de frecuencia de siniestros usando los GLM aplicables. Este modelo permitirá obtener una prima base y clases de riesgo basados en **características** heterogéneas, con el objetivo que cada grupo de riesgo se conforme de pólizas de asegurados con **características** homogéneas de comportamiento y riesgo.

Por otro lado, cabe resaltar que el GLM seleccionado proporcionará información a priori clave al Sistema Bonus-Malus (SBM), **específicamente** la frecuencia o tasa de siniestros esperada y la importancia relativa o peso de cada clase de riesgo en toda la población analizada.

Por último, para construir el SBM se debe establecer primero un número de niveles, el asegurado cambiará entre niveles (se le bonificará o se le penalizará) en función al número de siniestros que reporte en el periodo de análisis. Una vez se cuente con los niveles definidos en el SBM propuesto, se determinará para cada nivel un porcentaje de bonificación o penalización, denominado relatividad. De esa manera, se podrá determinar el valor de la prima ajustada para cada asegurado tan solo con multiplicar la prima base con la relatividad asociada al nuevo nivel del SBM en donde se encuentre el asegurado luego de revisar su historial de siniestros.

1.4. Organización del trabajo

En el **Capítulo 2**, presentamos los conceptos previos al desarrollo de un sistema bonus malus enfocándonos en las distribuciones más utilizadas para frecuencia

de siniestros, su modelación con covariables mediante los modelos lineales generalizados (GLM), y las cadenas de Markov que se utilizarán para definir las reglas y probabilidades de transición del SBM propuesto.

En el **Capítulo 3** desarrollaremos el modelo propuesto para el Sistema Bonus-Malus (SBM).

En el **Capítulo 4** desarrollaremos la aplicación del modelo propuesto con datos reales referentes a un seguro vehicular.

Finalmente, en el **Capítulo 5** discutimos algunas conclusiones obtenidas de este trabajo.

En el anexo presentaremos los códigos utilizados en la aplicación.



Capítulo 2

Conceptos previos

Se dedica el presente capítulo a introducir los conceptos previos que permitirán en el siguiente capítulo modelar el Sistema Bonus Malus (SBM) propuesto. En ese sentido, de manera inicial, y orientando siempre nuestro trabajo a un seguro automovilístico, se presenta a continuación algunos conceptos asociados al SBM propuesto:

- Siniestro: Accidente asociado al vehículo asegurado, el cual está cubierto por la póliza de seguros contratada.
- Reclamación: Es el registro del siniestro reportado por el asegurado.
- Prima: Monto que debe pagar periódicamente un asegurado por la póliza de seguro contratada.
- Tarifación del asegurador: Sistema de primas por tipo de asegurado diseñado por la compañía de seguros.

En complemento a los conceptos anteriores relacionados al sector de seguros, es importante considerar la heterogeneidad de los riesgos asociados a cada asegurado, pues, no todos los asegurados tienen el mismo perfil de riesgo y por ende no deberían de pagar lo mismo. En ese sentido, es razonable considerar que, si una compañía aseguradora no vende pólizas con primas diferenciadas, es decir, según el riesgo específico de cada asegurado, entonces dicha compañía podría experimentar mayores dificultades para retener algunos de sus clientes o incluso captar nuevos, ocasionando pérdidas para la empresa. En esa línea, si consideramos que, para la empresa aseguradora son buenos clientes aquellos que no han presentado siniestros en el periodo contratado, y son clientes no tan atractivos (malos clientes) aquellos que han presentado uno o más siniestros; entonces un escenario posible es que algunos de los buenos clientes opten por irse a la competencia, si descubren que pagan lo mismo que los malos clientes, es decir, si descubren que la aseguradora no premia (o permite pagar menos) al asegurado que tiene un mejor perfil de riesgo que los otros. Esta heterogeneidad se puede representar como un parámetro de riesgo λ que describe las características de riesgo asociada a cada asegurado, tomando en cuenta ello, es común que las aseguradoras se esfuercen en recolectar la mayor cantidad de

datos posibles y disponibles sobre sus asegurados, debido a que permite caracterizarlos en términos de sus datos personales (edad, sexo, salario, etc.), datos de su **vehículo** (tipo o modelo, motor, costo, tipo de combustible usado, etc.), o incluso datos del propio seguro adquirido (parcial o todo riesgo), dichos datos son resumidos en un vector de información \mathbf{x} .

En relación al concepto de prima, esta se puede definir (Denuit M., 2007) de la siguiente manera:

$$\pi_{base} = E(N | \Lambda = \lambda) \quad (2.1)$$

donde N es el número de siniestros reportados por el asegurado, y λ es el parámetro de riesgo asociado a cada asegurado con **características** \mathbf{x} . Asimismo, supondremos que todo siniestro reportado genera un pago del seguro de una unidad monetaria (*u.m.*). Si bien es posible tratar de modelar el pago de la prima de un seguro en términos de *u.m.*, es decir, estimar las primas basadas en la severidad de los siniestros, restringiremos nuestro tratamiento en este trabajo solo a las frecuencias de los siniestros reportados por cada asegurado.

Por otro lado, es importante mencionar que el SBM propuesto considera:

- Información a priori: esta información permite identificar variables explicativas (edad, sexo, potencia del **vehículo**, antigüedad del **vehículo**, entre otros), las cuales se usan para establecer las clases de riesgo diferenciadas de pólizas de seguros de automóvil. Este esquema se conoce como clasificación a priori. Asimismo, cabe resaltar que, si bien este esquema no logra identificar todos los factores de riesgo de los asegurados, tales como la agresividad del conductor detrás del volante o la rapidez de sus reflejos para evitar un posible accidente; es un punto de partida para segmentar de manera inicial a los asegurados según su perfil de riesgo observable, y a su vez determinar una prima base o también conocida como prima a priori, la cual se podría utilizar, por ejemplo, para un nuevo asegurado sin historial o experiencia en el uso del seguro a contratar con la empresa aseguradora.
- Información a posteriori: esta información permite re-evaluar o ajustar el valor de la prima base para cada asegurado, dado que, toma en cuenta el historial de reclamaciones del asegurado tan pronto dicha información se encuentra disponible, usualmente un año después de contratado el seguro. Este esquema se denomina clasificación a posteriori, y se basa en que, el mejor predictor para el número de reclamaciones futuras de un asegurado está en función del número de reclamaciones reportadas en el pasado por ese mismo asegurado (Denuit M., 2007). Asimismo, cabe resaltar que, para utilizar este esquema se puede hacer uso de un sistema de credibilidad actuarial o el uso de un sistema basado en la experiencia de reclamaciones denominado comercialmente Sistema Bonus Malus (SBM).

Ambos tipos de información permitirán establecer un monto de prima para cada asegurado asociado a su perfil de riesgo.

Adicionalmente a los conceptos generales presentados al inicio del **capítulo**, a continuación, algunos conceptos técnicos asociados al SBM propuesto:

2.1. Modelos lineales generalizados (GLM)

Antes de definir que es un modelo lineal generalizado (GLM), es importante recordar que en un modelo lineal el valor de la variable dependiente Y_i , para una observación i ($i = 1, 2, \dots, n$), se modela como una función lineal de $(p - 1)$ variables independientes x_1, x_2, \dots, x_{p-1} de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i(p-1)} + s_i = \mathbf{x}_i^T \boldsymbol{\beta} + s_i,$$

donde s_i es un término de error tal que $E(s_i) = 0$.

Por otra parte, los GLM se pueden definir como una extensión de los modelos de regresión lineal y no lineal, que permite incorporar distribuciones de respuesta no normales. En un modelo lineal generalizado la distribución de la variable de respuesta sólo necesita ser un miembro de la familia exponencial de distribuciones que incluye, entre otras, la Normal, Poisson, Binomial negativa y Gamma, es decir, Y_i es una variable aleatoria con función de densidad de probabilidad exponencial. En ese sentido, los miembros de la mencionada familia exponencial tienen una función de probabilidad o densidad de la forma general

$$f(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\varphi}) = \exp \left[\frac{(\mathbf{y}\boldsymbol{\theta} - b(\boldsymbol{\theta}))}{a(\boldsymbol{\varphi})} + h(\mathbf{y}, \boldsymbol{\varphi}) \right],$$

donde $\boldsymbol{\varphi}$ es un parámetro de escala, $\boldsymbol{\theta}$ es el parámetro natural de la familia exponencial y a, b, h son funciones conocidas.

Adicionalmente, los GLM plantean un modelo para la media del valor de la variable de respuesta, la cual se expresa de la siguiente manera:

$$\boldsymbol{\mu}_i = E(Y_i) = \mathbf{g}^{-1}(\boldsymbol{\eta}_i) = \mathbf{g}^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}),$$

donde $\boldsymbol{\eta}_i$ es un predictor lineal y \mathbf{g} una función de enlace que relaciona el valor esperado de la respuesta observada con el predictor lineal.

Para fines del presente trabajo, nos centraremos en un GLM de respuesta de conteo entre los que destacan el modelo de regresión de Poisson y el modelo de regresión binomial negativo, esto con el objetivo de modelar nuestra variable de interés relacionada al número de siniestros reportado por un asegurado.

2.1.1. Regresión de Poisson

En este modelo, la variable respuesta Y_i es una variable aleatoria que sigue una distribución de Poisson con parámetro $\lambda_i > 0$ ($Y_i \sim \text{Poisson}(\lambda_i)$), es decir, Y_i tiene una función de probabilidad

$$P(Y_i = y | \lambda_i) = \frac{\lambda_i^y e^{-\lambda_i}}{y!}, \quad y = 0, 1, 2, \dots \quad (2.2)$$

con media y varianza:

$$E(Y_i) = \lambda_i \text{ y } \text{Var}(Y_i) = \lambda_i.$$

La expresión (2.2) puede también expresarse con notación de familia exponencial en su forma canónica

$$f(y, \theta, \varphi) = \exp \left\{ y \log(\lambda_i) - \lambda_i + \log \frac{1}{y!} \right\},$$

donde $h(y, \varphi) = \log \frac{1}{y!}$, el parámetro natural es $\theta = \log(\lambda_i)$, $a(\varphi) = 1$ y $b(\theta) = \lambda_i$.

Asimismo, el modelo de regresión de Poisson plantea que

$$\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i(p-1)},$$

donde se utiliza aquí una función de enlace logarítmica $g(\lambda_i) = \log(\lambda_i)$.

Equivalentemente, se tiene que:

$$\lambda_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}. \quad (2.3)$$

Adicionalmente, para estimar los parámetros de la regresión usaremos el método de máxima verosimilitud (Paula, 2013). En línea con ello, para una muestra aleatoria de n observaciones de la variable aleatoria Y_i , y los predictores \mathbf{x}_i , la función de verosimilitud es

$$\begin{aligned} L(\mathbf{y}, \boldsymbol{\beta}) &= \prod_{i=1}^n f(y_i | \lambda_i) = \prod_{i=1}^n \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \\ &= \frac{\prod_{i=1}^n \lambda_i^{y_i} \exp\left(-\sum_{i=1}^n \lambda_i\right)}{\prod_{i=1}^n y_i!}, \end{aligned}$$

donde $\lambda_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}$ y su función de log-verosimilitud es:

$$\log L(\mathbf{y}, \boldsymbol{\beta}) = \sum_{i=1}^n y_i \log(e^{\mathbf{x}_i^T \boldsymbol{\beta}}) - \sum_{i=1}^n e^{\mathbf{x}_i^T \boldsymbol{\beta}} - \sum_{i=1}^n \log(y_i!).$$

Se buscará maximizar esta función con respecto a $\boldsymbol{\beta}$ a través de métodos numéricos.

Por último, una vez obtenido los estimadores de máxima verosimilitud del parámetro, $\hat{\boldsymbol{\beta}}$, el modelo ajustado de regresión de Poisson será $\hat{\lambda}_i = e^{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}}$.

2.1.2. Regresión binomial negativa

En este modelo, la variable respuesta Y_i es una variable aleatoria que sigue una distribución binomial negativa, $Y_i \sim \mathbf{BN}(r, p_i)$, con parámetros $r > 0$ (real) y $p_i \in [0, 1]$, es decir, Y_i tiene una función de probabilidad:

$$P(Y_i = y | r, p_i) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p_i^r (1-p_i)^y, y = 0, 1, 2, \dots \quad (2.4)$$

con media y varianza:

$$E(Y_i) = \frac{r(1-p_i)}{p_i} \text{ y } Var(Y_i) = \frac{r(1-p_i)}{p_i^2}.$$

Adicionalmente, si re-parametrizamos esta distribución en términos de su media: $\mu_i = \frac{r(1-p_i)}{p_i}$ y r , se tendría la siguiente función de probabilidad:

$$P(Y_i = y | r, \mu_i) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \cdot \frac{r}{\mu_i+r} \cdot \left(1 - \frac{r}{\mu_i+r}\right)^y. \quad (2.5)$$

Por otra parte, la expresión (2.5) también puede expresarse con notación de una familia exponencial en su forma canónica

$$f(y | r, \mu_i) = \exp \left\{ y \log \left(1 - \frac{r}{\mu_i+r}\right) + r \log \frac{r}{\mu_i+r} + \log \frac{r}{\mu_i+r} \right\} \cdot \left(1 - \frac{r}{\mu_i+r}\right)^y,$$

con $h(y, \varphi) = \log \frac{r+y-1}{y}$, $a(\varphi) = y$, $b(\theta) = r \log \frac{r}{\mu_i+r}$ y parámetro natural $\theta = \log \left(1 - \frac{r}{\mu_i+r}\right)$.

Asimismo, el modelo de regresión binomial negativa plantea que

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i(p-1)}$$

donde se utiliza aquí nuevamente la función de enlace logarítmico $g(\mu_i) = \log(\mu_i)$.

Equivalentemente, se tiene que:

$$\mu_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}}. \quad (2.6)$$

Por otro lado, para estimar los parámetros de la regresión usaremos el método de máxima verosimilitud (Paula, 2013). En línea con ello, para una muestra aleatoria de n observaciones de la respuesta Y_i , y los predictores \mathbf{x}_i , la función de verosimilitud es

$$L(\mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^n f(y_i | r, p_i) = \prod_{i=1}^n \frac{\Gamma(y_i+r)}{\Gamma(y_i+1)\Gamma(r)} \cdot \frac{r}{\mu_i+r} \cdot \left(1 - \frac{r}{\mu_i+r}\right)^{y_i},$$

donde $\mu_i = e^{x_i^T \beta}$ y su log-verosimilitud es:

$$\log L(\mathbf{y}, \beta) = \sum_{i=1}^n \log \frac{y_i - 1}{r - 1} + r \log(r) - \sum_{i=1}^n y_i \log(e^{x_i^T \beta}) + \sum_{i=1}^n (y_i - r) \log(e^{x_i^T \beta} - r).$$

Se buscará maximizar esta función con respecto a β a través de métodos numéricos. Una vez obtenidos los estimadores de máxima verosimilitud del parámetro, $\hat{\beta}$, el modelo ajustado de regresión de binomial negativo será:

$$\hat{\mu}_i = e^{x_i^T \hat{\beta}}.$$

Finalmente, un resultado interesante plantea que el modelo binomial negativo puede verse como la mixtura de una distribución de Poisson y una distribución Gamma, según el cual la variable respuesta Y sigue una distribución Poisson, cuya media está especificada de forma incompleta debido a una situación de heterogeneidad no observada. Esto puede verificarse según el siguiente resultado:

Proposición 2.1.1.

Si $Y | \Delta \sim \text{Poisson}(\lambda \delta)$, donde $\lambda > 0$ y $\Delta \sim \Gamma(\tau, \omega)$, entonces $Y \sim \text{BN}(\tau, \frac{\omega}{\lambda + \omega})$.

Demostración:

De forma general

$$\begin{aligned} P(Y = y) &= \int_0^\infty P(Y = y | \Delta = \delta) f_\Delta(\delta) d\delta \\ P(Y = y) &= \int_0^\infty \frac{e^{-\lambda \delta} (\lambda \delta)^y}{y!} \cdot \frac{\omega^\tau}{\Gamma(\tau)} \delta^{\tau-1} e^{-\omega \delta} d\delta \\ P(Y = y) &= \frac{\lambda^y \omega^\tau}{y! \Gamma(\tau)} \int_0^\infty e^{-\theta(\lambda + \omega)} \delta^{(y + \tau) - 1} d\delta \end{aligned}$$

La integral última corresponde a la función Gamma de forma:

$$\int_0^\infty e^{-ax} x^{b-1} dx = \frac{\Gamma(b)}{a^b}$$

donde: $a = (\lambda + \omega)$ y $b = (y + \tau)$. Así $\int_0^\infty e^{-\delta(\lambda + \omega)} \delta^{(y + \tau) - 1} d\delta = \frac{\Gamma(y + \tau)}{(\lambda + \omega)^{y + \tau}}$

En consecuencia,

$$\begin{aligned} P(Y = y) &= \frac{\lambda^y \omega^\tau}{y! \Gamma(\tau)} \frac{\Gamma(y + \tau)}{(\lambda + \omega)^{y + \tau}} \\ P(Y = y) &= \frac{\Gamma(y + \tau)}{\Gamma(y + 1) \Gamma(\tau)} \frac{\lambda^y \omega^\tau}{(\lambda + \omega)^{y + \tau}} \\ P(Y = y) &= \frac{\Gamma(y + \tau)}{\Gamma(y + 1) \Gamma(\tau)} \frac{\omega}{\lambda + \omega} \left(1 - \frac{\omega}{\lambda + \omega}\right)^y \end{aligned}$$

y por tanto:

$$Y \sim \text{BN} \left(\tau, \frac{\omega}{\lambda + \omega} \right).$$

En particular si $Y | \Delta = \sum \delta \sim \text{Poisson}(\lambda \delta)$ y $\Delta \sim \Gamma(a, a)$ (de modo que $E(\Delta) = 1$), entonces $Y \sim \text{BN} \left(a, \frac{a}{\lambda + a} \right)$.

2.1.3. Regresión de Poisson inflacionada en ceros

En 1992, Lambert (1992) propuso el modelo de Poisson inflacionado en ceros, el cual postula de que existe una probabilidad π_i de que la i -ésima observación sea igual a cero. Es decir, este modelo postula que la variable respuesta Y_i tiene como función de probabilidad a:

$$P(Y_i = y | \pi_i, \lambda_i) = \begin{cases} \pi_i + (1 - \pi_i) e^{-\lambda_i} & \text{si } y = 0 \\ (1 - \pi_i) \frac{\lambda_i^y e^{-\lambda_i}}{y!} & \text{si } y > 0. \end{cases}$$

Asimismo, el modelo de regresión de Poisson inflacionada en ceros plantea que

$$\begin{aligned} \log(\lambda_i) &= \mathbf{z}_i^T \boldsymbol{\beta}, \\ \text{logit}(\pi_i) &= \log \frac{\pi_i}{1 - \pi_i} = \boldsymbol{\omega}_i^T \boldsymbol{\gamma}. \end{aligned}$$

Equivalentemente se tiene que

$$\begin{aligned} \lambda_i &= e^{\mathbf{z}_i^T \boldsymbol{\beta}} \\ \pi_i &= \frac{e^{\boldsymbol{\omega}_i^T \boldsymbol{\gamma}}}{1 + e^{\boldsymbol{\omega}_i^T \boldsymbol{\gamma}}}, \end{aligned} \quad (2.7)$$

siendo \mathbf{z}_i y $\boldsymbol{\omega}_i$ vectores de variables regresoras (las cuales pueden incluir o no términos en común); y $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ vectores de parámetros desconocidos.

Por otro lado, para una muestra aleatoria de n observaciones de la respuesta Y_i y los regresores $\mathbf{z}_i, \boldsymbol{\omega}_i$, la función de verosimilitud es

$$\begin{aligned} L(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{i=1}^n \log(\pi_i) + \sum_{y_i=0} \log\{e^{\boldsymbol{\omega}_i^T \boldsymbol{\gamma}} + \exp(-e^{\mathbf{z}_i^T \boldsymbol{\beta}})\} \\ &+ \sum_{y_i>0} (y_i \mathbf{z}_i^T \boldsymbol{\beta} - e^{\mathbf{z}_i^T \boldsymbol{\beta}}) - \sum_{y_i>0} \log(y_i!). \end{aligned}$$

Por último, se deberá maximizar el logaritmo de la función de verosimilitud con la finalidad de obtener los estimadores de máxima verosimilitud de los vectores de parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$.

2.1.4. Regresión binomial negativa inflacionada en ceros

Una distribución binomial negativa inflacionada en ceros es un resultado de mezclar una distribución de Bernoulli y una distribución binomial negativa, la cual según

Monzon (2009) está dada por:

$$P(Y_i = y | \pi_i, \mu_i, r) = \begin{cases} \pi_i + (1 - \pi_i)g(0) & \text{si } y = 0 \\ (1 - \pi_i)g(y) & \text{si } y > 0, \end{cases}$$

siendo g la función de probabilidad de una variable aleatoria binomial negativa reparametrizada en términos de su media $\mu_i = \frac{r(1-\pi_i)}{\pi_i}$ y r ; es decir, con función de densidad:

$$g(y) = P(Y_i = y | \mu_i, r) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \cdot \frac{r}{\mu_i + r} \left(1 - \frac{r}{\mu_i + r}\right)^y \quad (2.8)$$

El modelo de regresión binomial negativo inflacionado en ceros plantea que

$$\begin{aligned} \log(\mu_i) &= z_i^T \beta, \\ \text{logit}(\pi_i) &= \log \frac{\pi_i}{1 - \pi_i} = \omega_i^T \gamma. \end{aligned}$$

Equivalentemente se tiene que

$$\begin{aligned} \mu_i &= e^{z_i^T \beta} \\ \pi_i &= \frac{e^{\omega_i^T \gamma}}{1 + e^{\omega_i^T \gamma}}, \end{aligned} \quad (2.9)$$

siendo z_i y ω_i vectores de variables regresoras (las cuales pueden incluir o no términos en común), y β y γ vectores de parámetros desconocidos.

Para este caso, igual que con la regresión Poisson inflacionada en ceros, se estimarán los vectores de parámetros β y γ por máxima verosimilitud y métodos numéricos de aproximación (Paula, 2013).

2.2. Herramientas de diagnóstico y bondad de ajuste

A continuación, se muestran algunos de los test o pruebas que nos permitirán comparar modelos y diagnosticar si las variables y modelos seleccionados son los adecuados para los datos disponibles (bondad de ajuste) en el contexto de los seguros de automovil y el SBM propuesto.

2.2.1. Devianza

Es una medida de bondad de ajuste que compara el modelo saturado (considera todas las variables explicativas) con el modelo de interés (considera solo a un subconjunto de las variables explicativas). Este concepto se denomina devianza, la cual según Denuit M. (2007) se define como:

$$D = 2(L(\hat{\theta}, y) - L(\theta^*, y))$$

donde $\hat{\theta}$ es el estimador de máxima verosimilitud del vector de parámetros θ del

modelo saturado y $\hat{\theta}^*$ el estimador de máxima verosimilitud del vector de parámetros θ^* del modelo de interés. Además $L(\hat{\theta}, \mathbf{y})$ es la función log-verosimilitud para el modelo saturado y $L(\hat{\theta}^*, \mathbf{y})$ es la función log-verosimilitud para el modelo de interés .

Asimismo, si el modelo de interés es adecuado, la devianza tendrá una distribución asintótica ji-cuadrado χ^2 a medida que aumenta n , con los grados de libertad igual al número de observaciones menos el número de parámetros del modelo de interés.

Por otro lado, cabe resaltar que el grado de aproximación de la devianza a la distribución Ji-cuadrado depende del tipo de distribución que se trate.

2.2.2. Test de Equidispersión

La igualdad de la media y la varianza se conoce como la propiedad de equidispersión de Poisson. Esta propiedad es frecuentemente violada en datos reales (Cameron y Trivedi, 1998) dado que usualmente se encuentran casos en donde la varianza excede o es menor a la media pero no igual. Cuando la varianza excede a la media se le conoce como sobredispersión, en cambio cuando la varianza es menor que la media se le denomina subdispersión. Con base en la bibliografía señalada, cabe resaltar que la sobredispersión aparece con más frecuencia que la subdispersión.

En línea con el párrafo anterior y considerando la restricción de Poisson (igualdad de media y varianza), Cameron y Trivedi (1998) propone una prueba donde contrasta la hipótesis de equidispersión ($E(\mathbf{y}) = Var(\mathbf{y}) = \mathbf{u}$) contra la alternativa de que la varianza excede la media, es decir, $Var(\mathbf{y}) = \mathbf{u} + \mathbf{c}\mathbf{u}$, donde \mathbf{c} es una constante. Siendo la hipótesis nula $H_0 : \mathbf{c} = 0$ y la hipótesis alternativa $H_a : \mathbf{c} > 0$.

Asimismo, y en relación a las distribuciones revisadas en las secciones anteriores, si bien en el análisis de datos de conteo, el punto de partida es la distribución de Poisson, en la práctica la mayoría de veces resulta inapropiada debido a la sobredispersión. En ese sentido, a menudo la heterogeneidad no observada causará que la varianza exceda la media (fenómeno denominado sobredispersión), lo cual va en contra de la fuerte restricción de equidispersión que impone un modelo de regresión de Poisson. De otro lado, y a diferencia de la Poisson, la distribución binomial negativa es más flexible y puede soportar la sobredispersión presente en los datos.

2.2.3. Test de Vuong

Es una extensión de la prueba de razón de verosimilitud para evaluar modelos no anidados (por ejemplo, el modelo de regresión de Poisson y el modelo de regresión de Poisson inflacionado en ceros), la cual usa el criterio de información de Kullback-Leibler (KLIC), con el propósito de medir la cercanía de un modelo al verdadero. Por ello, si la distancia entre un modelo de regresión específico y el modelo verda-

dero se define como el **mínimo** valor de KLIC sobre las distribuciones en el modelo, entonces es natural definir el mejor modelo de regresión entre una colección de modelos competidores como el modelo más cercano a la verdadera distribución. [Vuong \(1989\)](#) propone para el caso de dos modelos competitivos un test de mayor **cercanía** de uno de estos modelos al verdadero.

2.2.4. AIC

Es una medida de la calidad relativa de un modelo **estadístico** dado un conjunto de datos, la cual utiliza la estimación de probabilidad máxima y el número de parámetros k para estimar la información no capturada en el modelo. Este concepto se deriva de la **teoría** de la información y está diseñado para elegir el modelo que produce una distribución de probabilidad con la menor discrepancia de la distribución verdadera (medida por la discrepancia de Kuhlbeck-Liebner ([Bozdogan, 2000](#))).

El AIC se define por:

$$AIC = 2k - 2\log(\hat{L}),$$

donde k es el número de parámetros en el modelo y \hat{L} es el máximo valor de la función verosimilitud para el modelo estimado. Un valor AIC más bajo indica menos información perdida, por lo tanto, un modelo que predice mejor.

Por último, cabe resaltar que el AIC no sólo ofrece una compensación a la exactitud sino que también penaliza la complejidad del modelo y además busca seleccionar el modelo que proporcione las mejores predicciones, dado un conjunto de modelos candidatos.

2.2.5. BIC

Es una medida de bondad de ajuste para la selección de modelos entre un conjunto finito de estos. Este concepto se basa en la función de probabilidad, y está estrechamente relacionado con el criterio de información de Akaike (AIC). Asimismo, el BIC introduce un término de penalización más severo por el número de parámetros estimados en el modelo, por lo que tiende a elegir modelos con menor número de variables explicativas es decir modelos más parsimoniosos. Su formulación viene dada por:

$$BIC = k\log(n) - 2\log(\hat{L}),$$

donde n es el número de observaciones o tamaño de muestra, k es el número de parámetros en el modelo **estadístico** y \hat{L} es el máximo valor de la función verosimilitud para el modelo estimado.

2.3. Cadenas de Markov

Todo Sistema Bonus-Malus (SBM) está definido por una cadena de Markov con una cierta cantidad de niveles o escalas de pago y una matriz de transición entre sus niveles.

Una cadena de Markov de tiempo y estados discretos es un proceso estocástico $\{X_t\}$ en el que una variable aleatoria X_t va **cambiando con el paso del tiempo 't'**. Asimismo, las cadenas de Markov tienen la propiedad de que la probabilidad de que $X_t = j$ sólo depende del estado inmediatamente anterior del sistema: X_{t-1} , siendo el espacio de estados denotado por $S = \{1, 2, \dots, s\}$, el cual consideraremos en este trabajo, finito y con 's' estados.

Adicionalmente, cuando en una cadena de Markov dichas probabilidades no dependen del tiempo 't' en que se considere; es decir que

$$P(X_t = j | X_{t-1} = i) = p_{i,j}, \quad \forall i, j \in S,$$

entonces la cadena se dice homogénea y a la probabilidad $p_{i,j}$ se le denomina la probabilidad de transición del estado i en el tiempo $t-1$ al estado j en el tiempo t .

Estos satisfacen la condición $p_{i,j} > 0$ y $\sum_{j=1}^s p_{i,j} = 1$, para cada $i = 1, 2, \dots, s$.

Para cada i fijo, la serie de valores $(p_{i,j})_{j \in S}$, conforman una distribución de probabilidades. Todos estos valores se combinan formando una matriz de transición P de tamaño $s \times s$, la cual se expresa de la siguiente manera:

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,s} \\ p_{2,1} & p_{2,2} & \dots & p_{2,s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s,1} & p_{s,2} & \dots & p_{s,s} \end{pmatrix}$$

Por otra parte, la probabilidad de que la cadena esté en el estado j después de n pasos, dado que la cadena empezó en el estado i es expresada por la probabilidad de transición de orden n siguiente:

$$p_{i,j}^{(n)} = P(X_n = j | X_0 = i)$$

Note que estas probabilidades satisfacen que:

$$\begin{aligned} \sum_{j=1}^s p_{i,j}^{(n)} &= P(X_n = j, X_0 = i) \\ &= \sum_{k=1}^s P(X_n = j | X_{n-1} = k, X_0 = i) P(X_{n-1} = k | X_0 = i) \\ &= \sum_{k=1}^s p_{i,k}^{(n-1)} p_{k,j} \end{aligned}$$

de donde se puede deducir que:

$$P^{(n)} = \begin{pmatrix} p_{1,1}^{(n)} & p_{1,2}^{(n)} & \dots & p_{1,s}^{(n)} \\ p_{2,1}^{(n)} & p_{2,2}^{(n)} & \dots & p_{2,s}^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s,1}^{(n)} & p_{s,2}^{(n)} & \dots & p_{s,s}^{(n)} \end{pmatrix} = P^{(n-1)}P = P^{(n-2)}P^2 = \dots = P^n$$

Definición 2.1. *Distribución estacionaria*

Sea X_n una cadena de Markov con espacio de estados S y matriz de transición P . Una distribución $\pi = (\pi_1, \pi_2, \dots, \pi_i, \dots, \pi_s)$ se dice que es estacionaria o de equilibrio si

$$\pi = \pi P.$$

Para calcular la distribución estacionaria, se debe resolver el siguiente sistema:

$$\begin{aligned} \pi^T &= \pi^T P \\ \pi e &= 1, \end{aligned}$$

donde e es un vector columna de unos de orden s . Sea E la matriz de unos de orden $s \times s$. De ello, según Rolski (1999), se desprende que:

$$\pi^T = e^T (I - P + E)^{-1}.$$

Definición 2.2. *Cadena irreducible*

Una cadena irreducible es aquella en la que todos los estados son alcanzables desde cualquier otro estado de la cadena en un número finito de pasos. Eso implica que se puede llegar a cualquier estado j desde cualquier otro estado i , o más formalmente, si $p_{i,j}^{(n)} > 0$, para algún número entero n .

Definición 2.3. *Cadena aperiódica*

Se define el periodo de un estado i como $d(i) = \text{mcd } n | p_{i,i}^{(n)} > 0$. Este es el máximo común divisor (*mcd*) del conjunto de los enteros n para los que $p_{i,i}^{(n)} > 0$. Decimos que una cadena es aperiódica si para algún estado i se cumple que $d(i) = 1$.

Teorema 2.1.

Dada la matriz de transición P de una cadena de Markov finita, aperiódica e irreducible con s estados, existe una única solución al sistema de ecuaciones

$$\pi_j = \sum_{i=1}^s \pi_i p_{i,j}, \quad \forall i, j = 1, \dots, s$$

tal que $\sum_{i=1}^s \pi_i = 1$.

Además, esta solución viene dada por

$$\pi_j = \lim_{n \rightarrow \infty} p_{i,j}^{(n)}, \quad \forall i = 1, \dots, s.$$

Demostración: ver [Haggstom \(2002\)](#)

El teorema presentado postula que, para n suficientemente grande, sea cual fuese la distribución inicial, la distribución de probabilidad de X_n convergerá a la distribución estacionaria π .



Capítulo 3

Sistema Bonus-Malus

Los Sistemas Bonus Malus para seguros de automóvil permiten premiar a los asegurados que no presentan siniestros y castigarlos en caso contrario. Es decir, un asegurado recibirá un descuento del monto de su prima base por no presentar siniestros, y en otro caso, otro asegurado recibirá un aumento en el monto de su prima base por presentar uno o más siniestros en el periodo analizado. Este proceso se realiza por lo general anualmente, y el nuevo monto de prima a pagar (prima ajustada) por el asegurado aplica desde el año siguiente. Este premio (descuento en la prima) o castigo (aumento en la prima) se da en términos de transiciones entre escalas de pago. En general, en los seguros de automóvil la mayoría de los asegurados no presentan siniestros y por esto es necesario que las aseguradoras usen métodos y sistemas creativos para ajustar a favor del cliente los montos de las primas de todos aquellos asegurados que no presentan accidentes (siniestros) con el objetivo de asegurar su retención y fidelización. Uno de esos métodos creativos usados en la actualidad se denomina Sistema Bonus Malus (SBM). Este SBM incorpora el historial de siniestros de un asegurado en un año (información a posteriori) y actualiza el valor de la prima que pagó este asegurado al iniciar el año (prima base) con un factor de ajuste denominado relatividad, el resultado obtenido de la multiplicación de la prima base con la relatividad se denomina prima ajustada.

Por otro lado, otra **característica** de los sistemas bonus malus es que generan tarifas de primas únicas para cada uno de los niveles del SBM. Esto simplifica y facilita la gestión de tarifarios y preciarlos tanto para el asegurado (consumidor final) como para la fuerza de ventas de la empresa aseguradora dado que permite agilizar la explicación de cuánto el asegurado pagará por el seguro contratado.

El Sistema Bonus Malus (SBM) propuesto se resume en la siguiente expresión:

$$\pi_{ajustada} = \pi_{base} r_A \quad (3.1)$$

donde:

- La prima ajustada ($\pi_{ajustada}$) representa la nueva prima considerando el historial de reclamaciones del asegurado, y corresponde al producto de las variables π_{base} y r_A .
- La prima base (π_{base}) corresponde al monto que determina la aseguradora al

inicio de la relación contractual con el asegurado.

- La relatividad (r_A) corresponde al porcentaje de bonificación (descuento) o penalización (incremento) obtenido luego de aplicar las reglas del SBM propuesto para el nivel A .

El SBM propuesto se basa en dos procedimientos; una clasificación a priori y una clasificación a posteriori. A continuación se detalla cada uno de ellos.

3.1. Clasificación a priori

La clasificación a priori permite segmentar la cartera de clientes en clases de riesgo (lo más homogéneos posible) para poder determinar el monto de prima que pagará cada asegurado (prima base) según el perfil de riesgo en donde se encuentre. En el SBM propuesto la clasificación a priori se basará en el uso de los GLM con respuesta de conteo que se detallaron en el **Capítulo 2**, específicamente, se hará uso del GLM de mejor ajuste luego de aplicar todos los criterios de comparación entre los modelos de regresión analizados (sección 2.2: herramientas de diagnóstico y bondad de ajuste). En ese sentido, la prima base se podría calcular con modelos de regresión de Poisson, binomial negativa, así como modelos de Poisson y binomial negativo inflacionado en ceros, que son los modelos de conteo que consideraremos en este trabajo. Asimismo, en la sección de aplicación, se considerará como variable respuesta a el número de siniestros asociado al i -ésimo asegurado durante el periodo de un año y se denotará como N_i , el cual se estimará **vía** un modelo de regresión por máxima verosimilitud.

De otro lado, dada la complejidad de las ecuaciones de verosimilitud obtenidas, se deberá recurrir a procedimientos numéricos para su solución tales como el método de Newton-Raphson o el método Fisher scoring (Cordeiro G.M., 2000).

A partir de los parámetros estimados del modelo ajustado se podrá obtener el número esperado de siniestros $E(N_i)$ para el i -ésimo asegurado, lo cual es lo más relevante para la clasificación a priori en el presente trabajo.

Por tanto, a partir de los resultados del modelo ajustado y haciendo uso de las covariables más significativas, se puede clasificar a priori a los asegurados en un número k de clases o perfiles de riesgo, en donde por cada clase se identifica su frecuencia esperada de siniestros denotada por λ_k y un peso asociado a dicha clase de riesgo denotado por ω_k . Estas dos variables λ_k y ω_k alimentarán el SBM propuesto de modo que sea posible indicar que el SBM incorpora información a priori.

3.2. Clasificación a posteriori

El esquema de clasificación a posteriori toma como entrada de información los resultados de la clasificación a priori con el objetivo de actualizar la prima base considerando un factor de corrección denominado relatividad r_A , este factor significa que el titular de la póliza que ocupa un nivel A en la escala Bonus Malus debe pagar r_A veces la prima base, este resultado se denomina prima ajustada. Es por ello, que en el contexto de un SBM se indica que existe una escala de pagos diferenciada, es

decir, un monto distinto para cada uno de los niveles del SBM analizado. Asimismo, es importante indicar que la transición entre niveles A en dicha escala toma en cuenta la frecuencia de siniestros reportados por cada asegurado. En ese sentido, en el SBM propuesto la clasificación a posteriori se basa en la probabilidad de que un asegurado se encuentre en un determinado nivel A del Sistema Bonus Malus (SBM).

Por otra parte, si bien para los casos de nuevos asegurados (de los cuales no se cuenta con un historial de siniestros reportados), se hace uso de la clasificación a priori, es relevante indicar que no es posible observar todos los factores de riesgo importantes únicamente con características a priori. En ese sentido, tan pronto se encuentre disponible el historial de siniestros del asegurado de interés, esta información deberá servir para ajustar la prima base, dado que, dicha información revela el nivel de riesgo materializado de cada conductor o asegurado.

Uno de los métodos que existen para ajustar la prima a posteriori ($\pi_{ajustada}$) es la clasificación basada en la experiencia del asegurado o SBM, el cual penaliza las reclamaciones del conductor y premia el historial impecable del mismo. Es importante indicar que se deberá trabajar con datos transversales si se trata de información de un solo periodo (por ejemplo, el resumen de un año), en cambio, en caso se cuente con información de varios periodos (por ejemplo, varios años de historial de siniestros de un grupo de asegurados), se deberá trabajar con datos longitudinales o panel. En el primer caso, se hará uso de los GLM (usados también para la clasificación a priori) y en el segundo caso, se deberá hacer uso de los GLMM (modelos lineales generalizados mixtos). No obstante, el presente trabajo solo considerará el sistema de tarificación a posteriori basado en los modelos GLM. Para mayor detalle sobre el modelo GLMM, ver [Antonio y Van Berkum \(2015\)](#).

3.2.1. Componentes principales de un SBM

Según [Denuit M. \(2007\)](#) todo SBM cuenta con los siguientes componentes:

- Una prima base (π_{base}): correspondiente al valor pagado por el asegurado inicialmente (al contratar el seguro). También denominada prima a priori dado que se hace uso del GLM para estimar el monto de prima base diferenciada que debe pagar cada asegurado de acuerdo a su clase de riesgo.
- Una relatividad (r_A): correspondiente al factor porcentual que indica cuantas veces tendrá que pagar el asegurado el valor de la prima base. El cálculo de la relatividad para el SBM propuesto se detalla en la sección 3.5.
- Una prima a posteriori o ajustada ($\pi_{ajustada}$): correspondiente al producto de las variables r_A y π_{base} , este producto corresponde al valor que pagará el asegurado como resultado de la aplicación de las nuevas tarifas de la compañía de seguros aplicando el SBM.
- Una escala Bonus-Malus, conformada por A niveles, numerados desde el 0 hasta $A-1$, considerando el nivel 0 como la escala más baja de pago. El asegurado se

encontrará, inicialmente, en alguna escala del SBM. Posteriormente, el asegurado se mantendrá en su mismo nivel o pasará a otro nivel de la escala según las reglas de transición definidas en el SBM propuesto. La escala Bonus-Malus propuesta para el presente trabajo se detalla en la sección 3.3.

- Reglas y probabilidades de transición, las cuales indican como un asegurado transita de un nivel a otro según su historial de siniestros en un periodo evaluado con una probabilidad determinada (en la cual se utiliza la **teoría** de Cadenas de Markov). Las reglas y probabilidades de transición del SBM propuesto se detallan en la sección 3.4.

3.2.2. Modelo de credibilidad del SBM propuesto

Para la obtención de la prima ajustada ($\pi_{ajustada}$) de un asegurado en nuestro SBM, se asumirá un modelo de credibilidad como el descrito en [Antonio y Van Berkum \(2015\)](#), el cual presenta las siguientes **características**:

- **Se introduce una variable aleatoria Λ cuyo** valor observado para el sujeto i se denotará por λ_i , de acuerdo a lo indicado en la ecuación (2.1). Este es un término de riesgo o también denominado de heterogeneidad observable, que se puede resumir y extraer de la información recolectada por la aseguradora, por variables asociadas a los datos personales de los asegurados (edad, sexo, salario), datos de sus **vehículos** asegurados (tipo o modelo, motor, costo, tipo de combustible usado), o incluso los datos del propio seguro adquirido (cobertura parcial o a todo riesgo), entre otros datos.
- Se define un **'parámetro'** de heterogeneidad Θ no observable que representa la información latente que no es posible de recolectar ni recuperar por la aseguradora, como por ejemplo, los reflejos del conductor, su estilo de conducción o su experiencia de manejo, entre otros. Según [Alcaide \(2015\)](#), este parámetro aparece como resultado de incorporar un término de perturbación (s_i) en el modelo de regresión de Poisson

$$\mu_i^* = e^{(x_i^T \beta + s_i)} = e^{x_i^T \beta} e^{s_i}.$$

Luego, considerando $\lambda_i = e^{x_i^T \beta}$ (ver (2.2)), es posible reescribir la expresión anterior como:

$$\mu_i^* = \lambda_i \Theta_i,$$

donde la perturbación $\Theta_i = e^{s_i}$ representa la heterogeneidad no observable Θ_i . En cambio, λ_i corresponde a un parámetro de riesgo o heterogeneidad que sí es posible observar en función a las covariables disponibles en el modelo analizado.

- En el sistema todo asegurado i del portafolio será representado por el par (Θ_i, N_i) , donde Θ_i , como se indicó en el punto anterior, es un parámetro que representará la heterogeneidad no explicada para este asegurado y $N_i = (N_{i1}, N_{i2}, \dots)$ es su frecuencia de siniestros reportada. Asimismo, consideraremos que los pares (Θ_i, N_i) son independientes para cada asegurado i según el modelo de credibilidad de [Antonio y Van Berkum \(2015\)](#).
- Dado $\Theta_i = \theta$, las variables aleatorias N_i se asumirán independientes y con distribución de Poisson de parámetro $\vartheta = \lambda_i \theta$. Asimismo, asumiremos que Θ_i tiene una distribución Gamma con el mismo parámetro de forma y escala (a, a) , lo cual se hace para asegurar que $E(\Theta_i) = 1$, condición que debe cumplirse según el enfoque para el cálculo de las relatividades de [Norberg \(1976\)](#).
- Luego, de acuerdo a la proposición 2.1.1, dado que $N_i | \Theta_i = \theta \sim \text{Poisson}(\lambda_i \theta)$ y $\Theta_i \sim \Gamma(a, a)$, se cumple que

$$N_i \sim \text{BN} \left(a, \frac{a}{a + \lambda_i} \right) \quad (3.2)$$

Así, $E(N_i) = E(E(N_i | \Theta_i)) = \lambda_i E(\Theta_i) = \lambda_i$, de modo que λ_i se puede estimar utilizando un modelo de regresión generalizado, tal como se vió en el capítulo 2.

3.3. Escala del Sistema Bonus-Malus propuesto

La escala SBM propuesta en este trabajo estará conformada por 6 niveles, los cuales son numerados desde el 0 hasta el 5. Es importante resaltar que el número de niveles de una escala de un sistema Bonus Malus es diversa y **varía de país en país**, por ejemplo, existen casos emblemáticos de SBM en Bélgica (6 niveles), Reino Unido (7 niveles) e Irlanda (6 niveles). Un compendio referencial se encuentra en [Lemaire \(1995\)](#). Asimismo, cabe mencionar que el trabajar con un número de niveles **específico** implica a la vez desarrollar unas reglas de transición y probabilidades de transición determinadas, así como su respectiva distribución estacionaria, es decir, modificar el número de niveles de la escala de un SBM modifica también todos los conceptos anteriores asociados.

Esta escala permite definir niveles diferenciados para realizar la tarificación. Por ello, para cada nivel (A) en la escala del SBM, existirá una relatividad asociada r_A , la cual determina si la prima a priori (prima base) del conductor asegurado será descontada o penalizada. En función al número de siniestros reportados durante un periodo, el asegurado se moverá en la escala de niveles hacia arriba o hacia abajo. La escala SBM propuesta en este trabajo es conocida en la **bibliografía** revisada en [Antonio y Van Berkum \(2015\)](#) como “ $-1/\text{topscale}$ ”. Ver imagen inferior como referencia.

En esta escala un periodo sin reclamaciones resultará en una bonificación, lo que implica que el asegurado descienda un nivel, caso contrario, con una o más reclama-

| Niveles SBM | Descripción del nivel |
|-------------|--|
| 0 | Nivel más bajo de la escala del SBM (menor pago) |
| 1 | ↑ |
| 2 | |
| 3 | |
| 4 | |
| 5 | Nivel más alto de la escala del SBM (mayor pago) |

Figura 3.1: Escala “-1/topscale” del SBM propuesto

ciones provocará una penalización, que en el caso del SBM propuesto significa que el asegurado irá directamente al nivel 5, sin importar en qué nivel estuvo antes.

3.4. Reglas de transición, probabilidades de transición y distribución estacionaria

El estado o nivel del asegurado en el periodo t se denotará como $L_t(\vartheta)$ para enfatizar la dependencia de la frecuencia anual de su número de siniestros esperado ϑ . Sin embargo, el argumento ϑ en $L_t(\vartheta)$ no significa que los $L_t(\vartheta)$'s son funciones del parámetro ϑ , sino que su distribución depende de ϑ para un asegurado con frecuencia de siniestro esperado ϑ .

De manera más formal, la trayectoria del asegurado en la escala Bonus Malus se modela mediante una cadena de Markov $\{L_t(\vartheta)\}$, $t = 0, 1, 2, 3, \dots$, de variables aleatorias con espacio de estados $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$, tal que $L_t(\vartheta)$ representa el estado o nivel $\{0, 1, 2, 3, 4, 5\}$ ocupado durante el año, es decir, durante el intervalo de tiempo en años $[t, t + 1]$ por el asegurado. En ese sentido, una vez que el número N_t de los siniestros reportados por el asegurado se conoce, esta información se utiliza para reevaluar la posición del asegurado en la escala bonus-malus.

En los párrafos anteriores, se aprecia que la trayectoria futura del asegurado en la escala del SBM es solo dependiente del nivel actual (tiempo t) y es independiente de los niveles ocupados en el pasado ($t - 1, t - 2, t - 3, \dots$). Esta propiedad de independencia condicional es fundamental para aplicar los modelos de Markov.

Luego, podemos introducir el concepto de reglas de transición, el cual corresponde a la probabilidad de moverse de un nivel a otro en función del número de

siniestros reportados durante el año actual. Para ello, sea:

$$m_{i,j}(n) = \begin{cases} 1 & \text{, si la póliza cambia de nivel } i \text{ a nivel } j \\ 0 & \text{, en caso contrario} \end{cases}$$

donde $n \geq 0$ es el número de siniestros reportados; y escribamos estas reglas en la matriz $M(n)$:

$$M(n) = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{matrix} m_{0,0}(n) & m_{0,1}(n) & m_{0,2}(n) & m_{0,3}(n) & m_{0,4}(n) & m_{0,5}(n) \\ m_{1,0}(n) & m_{1,1}(n) & m_{1,2}(n) & m_{1,3}(n) & m_{1,4}(n) & m_{1,5}(n) \\ m_{2,0}(n) & m_{2,1}(n) & m_{2,2}(n) & m_{2,3}(n) & m_{2,4}(n) & m_{2,5}(n) \\ m_{3,0}(n) & m_{3,1}(n) & m_{3,2}(n) & m_{3,3}(n) & m_{3,4}(n) & m_{3,5}(n) \\ m_{4,0}(n) & m_{4,1}(n) & m_{4,2}(n) & m_{4,3}(n) & m_{4,4}(n) & m_{4,5}(n) \\ m_{5,0}(n) & m_{5,1}(n) & m_{5,2}(n) & m_{5,3}(n) & m_{5,4}(n) & m_{5,5}(n) \end{matrix} \end{matrix}$$

Entonces, $M(n)$ es una matriz de 1's y 0's donde cada fila de la matriz debe sumar 1 debido a que solo puede existir un estado o nivel al cual transite la cadena.

En el caso “-1/topscale”, por ejemplo

$$M(0) = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix} \end{matrix}, \quad (3.3)$$

pues si el asegurado no presentó siniestro alguno ($n = 0$) deberá transitar (bonificación) del estado s al estado $s - 1$ (para $s = 1, 2, 3, 4$ y 5). Es decir, el asegurado recibirá una bonificación por parte de las reglas del SBM, dado que, mientras el asegurado se encuentre en un nivel más cercano al $s = 0$ (primer estado del SBM propuesto) pagará menos que si se encuentra en un nivel más cercano al $s = 5$ (último estado del SBM propuesto).

De otro lado, si el asegurado presentó uno o más siniestros ($n > 0$) deberá transitar (penalización) directamente al estado $s = 5$ sin importar en qué nivel estuvo anteriormente, es decir, se tendrá:

$$M(1) = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{matrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix} \end{matrix} \quad (3.4)$$

y $M(n) = M(1)$ para todo $n = 2, 3, 4, 5, \dots$

De acuerdo al caso presentado anteriormente, las transiciones en el SBM “ $-1/topscale$ ” se aprecian en el Cuadro 3.1.

| Nivel inicial | Nivel ocupado si no hay siniestros reportados | Nivel ocupado si hay uno o más siniestros reportados |
|---------------|---|--|
| 0 | 0 | 5 |
| 1 | 0 | 5 |
| 2 | 1 | 5 |
| 3 | 2 | 5 |
| 4 | 3 | 5 |
| 5 | 4 | 5 |

Cuadro 3.1: Reglas de transición para la “ $-1/topscale$ ”

Por otra parte, asumiendo el modelo dado en (3.2), la probabilidad de transición de un asegurado del nivel A_1 al nivel A_2 en un paso viene dada en la cadena por:

$$\begin{aligned}
 p_{A_1, A_2}(\vartheta) &= P(L_{t+1}(\vartheta) = A_2 | L_t(\vartheta) = A_1) \\
 &= \sum_{n=0}^{\infty} P(L_{t+1}(\vartheta) = A_2 | N_{t+1} = n, L_t(\vartheta) = A_1) P(N_{t+1} = n | L_t(\vartheta) = A_1) \\
 &= \sum_{n=0}^{\infty} m_{A_1, A_2}(n) \frac{\vartheta^n \exp(-\vartheta)}{n!},
 \end{aligned} \tag{3.5}$$

con $A_1, A_2 \in \{0, 1, 2, 3, 4, 5\}$.

En la expresión anterior se hace uso del conocimiento previo sobre que $L_t(\vartheta)$ solo depende de N_1, N_2, \dots, N_t , es decir, depende del presente y del pasado pero no del futuro N_{t+1} . En ese sentido, se usará en el presente trabajo la independencia de N_{t+1} y $L_t(\vartheta)$ para presentar la matriz de probabilidad de transición de un solo paso, la cual viene dada en la cadena por:

$$P(\vartheta) = \begin{matrix} \square & & & & & & \\ & p_{0,0}(\vartheta) & p_{0,1}(\vartheta) & p_{0,2}(\vartheta) & p_{0,3}(\vartheta) & p_{0,4}(\vartheta) & p_{0,5}(\vartheta) \\ \square & p_{1,0}(\vartheta) & p_{1,1}(\vartheta) & p_{1,2}(\vartheta) & p_{1,3}(\vartheta) & p_{1,4}(\vartheta) & p_{1,5}(\vartheta) \\ \square & p_{2,0}(\vartheta) & p_{2,1}(\vartheta) & p_{2,2}(\vartheta) & p_{2,3}(\vartheta) & p_{2,4}(\vartheta) & p_{2,5}(\vartheta) \\ \square & p_{3,0}(\vartheta) & p_{3,1}(\vartheta) & p_{3,2}(\vartheta) & p_{3,3}(\vartheta) & p_{3,4}(\vartheta) & p_{3,5}(\vartheta) \square \\ \square & p_{4,0}(\vartheta) & p_{4,1}(\vartheta) & p_{4,2}(\vartheta) & p_{4,3}(\vartheta) & p_{4,4}(\vartheta) & p_{4,5}(\vartheta) \\ & p_{5,0}(\vartheta) & p_{5,1}(\vartheta) & p_{5,2}(\vartheta) & p_{5,3}(\vartheta) & p_{5,4}(\vartheta) & p_{5,5}(\vartheta) \end{matrix}$$

Específicamente, se aprecia que la probabilidad de transición para el caso de la “ $-1/topscale$ ” del SBM propuesto en este trabajo es

$$\begin{aligned}
p_{A_1, A_2}(\vartheta) &= \sum_{n=0}^{\infty} \frac{\vartheta^n}{n!} \exp(-\vartheta) m_{A_1, A_2}(n) \\
&= \frac{\vartheta^0}{0!} \exp(-\vartheta) m_{A_1, A_2}(0) + \frac{\vartheta^1}{1!} \exp(-\vartheta) m_{A_1, A_2}(1) + \frac{\vartheta^2}{2!} \exp(-\vartheta) m_{A_1, A_2}(2) \\
&\quad + \frac{\vartheta^3}{3!} \exp(-\vartheta) m_{A_1, A_2}(3) + \frac{\vartheta^4}{4!} \exp(-\vartheta) m_{A_1, A_2}(4) + \dots + \frac{\vartheta^n}{n!} \exp(-\vartheta) m_{A_1, A_2}(n).
\end{aligned}$$

Asimismo, usaremos las reglas de transición para la “-1/topscale” que aparecen en (3.3) y (3.4) considerando $m_{A_1, A_2}(n) = m_{A_1, A_2}(1)$ para $n \geq 2$ es decir

$$M(0) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \text{y} \quad M(1) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} = M(2) = \dots = M(n),$$

por ejemplo la expresión anterior para los casos $p_{0,0}(\vartheta)$ y $p_{0,1}(\vartheta)$ se desarrollan de la siguiente manera:

$$\begin{aligned}
p_{0,0}(\vartheta) &= \exp(-\vartheta) m_{0,0}(0) + \vartheta \exp(-\vartheta) m_{0,0}(1) + \frac{\vartheta^2}{2} \exp(-\vartheta) m_{0,0}(2) \\
&\quad + \frac{\vartheta^3}{6} \exp(-\vartheta) m_{0,0}(3) + \frac{\vartheta^4}{24} \exp(-\vartheta) m_{0,0}(4) + \dots + \frac{\vartheta^n}{n!} \exp(-\vartheta) m_{0,0}(n) \\
&= \exp(-\vartheta)(1) + \vartheta \exp(-\vartheta)(0) + \frac{\vartheta^2}{2} \exp(-\vartheta)(0) \\
&\quad + \frac{\vartheta^3}{6} \exp(-\vartheta)(0) + \frac{\vartheta^4}{24} \exp(-\vartheta)(0) + \dots + \frac{\vartheta^n}{n!} \exp(-\vartheta)(0) = \exp(-\vartheta)
\end{aligned}$$

$$\begin{aligned}
p_{0,1}(\vartheta) &= \exp(-\vartheta) m_{0,1}(0) + \vartheta \exp(-\vartheta) m_{0,1}(1) + \frac{\vartheta^2}{2} \exp(-\vartheta) m_{0,1}(2) \\
&\quad + \frac{\vartheta^3}{6} \exp(-\vartheta) m_{0,1}(3) + \frac{\vartheta^4}{24} \exp(-\vartheta) m_{0,1}(4) + \dots + \frac{\vartheta^n}{n!} \exp(-\vartheta) m_{0,1}(n) \\
&= \exp(-\vartheta)(0) + \vartheta \exp(-\vartheta)(0) + \frac{\vartheta^2}{2} \exp(-\vartheta)(0) \\
&\quad + \frac{\vartheta^3}{6} \exp(-\vartheta)(0) + \frac{\vartheta^4}{24} \exp(-\vartheta)(0) + \dots + \frac{\vartheta^n}{n!} \exp(-\vartheta)(0) = 0,
\end{aligned}$$

por tanto las probabilidades de transición $p_{A_1, A_2}(\vartheta)$ se muestran como elementos de la matriz $P(\vartheta)$

$$P(\vartheta) = \begin{matrix} \square & & & & & \square \\ \exp(-\vartheta) & 0 & 0 & 0 & 0 & 1 - \exp(-\vartheta) \\ \square \exp(-\vartheta) & 0 & 0 & 0 & 0 & 1 - \exp(-\vartheta) \square \\ 0 & \exp(-\vartheta) & 0 & 0 & 0 & 1 - \exp(-\vartheta) \\ 0 & 0 & \exp(-\vartheta) & 0 & 0 & 1 - \exp(-\vartheta) \square \\ \square 0 & 0 & 0 & \exp(-\vartheta) & 0 & 1 - \exp(-\vartheta) \square \\ 0 & 0 & 0 & 0 & \exp(-\vartheta) & 1 - \exp(-\vartheta) \end{matrix}$$

Consideremos ahora las potencias de P para analizar las probabilidades de transición de mayor orden

$$P(\vartheta)^2 = \begin{matrix} \square & & & & & \square \\ \exp(-2\vartheta) & 0 & 0 & 0 & -\exp(-2\vartheta) + \exp(-\vartheta) & 1 - \exp(-\vartheta) \\ \exp(-2\vartheta) & 0 & 0 & 0 & -\exp(-2\vartheta) + \exp(-\vartheta) & 1 - \exp(-\vartheta) \\ \square \exp(-2\vartheta) & 0 & 0 & 0 & -\exp(-2\vartheta) + \exp(-\vartheta) & 1 - \exp(-\vartheta) \square \\ 0 & \exp(-2\vartheta) & 0 & 0 & -\exp(-2\vartheta) + \exp(-\vartheta) & 1 - \exp(-\vartheta) \\ 0 & 0 & \exp(-2\vartheta) & 0 & -\exp(-2\vartheta) + \exp(-\vartheta) & 1 - \exp(-\vartheta) \\ 0 & 0 & 0 & \exp(-2\vartheta) & -\exp(-2\vartheta) + \exp(-\vartheta) & 1 - \exp(-\vartheta) \end{matrix}$$

$$P(\vartheta)^3 = \begin{matrix} \square & & & & & \square \\ \exp(-3\vartheta) & 0 & 0 & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \exp(-3\vartheta) & 0 & 0 & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \square \exp(-3\vartheta) & 0 & 0 & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \square \\ \exp(-3\vartheta) & 0 & 0 & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ 0 & \exp(-3\vartheta) & 0 & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ 0 & 0 & \exp(-3\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \end{matrix}$$

$$P(\vartheta)^4 = \begin{matrix} \square & & & & & \square \\ \exp(-4\vartheta) & 0 & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \exp(-4\vartheta) & 0 & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \square \exp(-4\vartheta) & 0 & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \square \\ \exp(-4\vartheta) & 0 & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \exp(-4\vartheta) & 0 & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ 0 & \exp(-4\vartheta) & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \end{matrix}$$

$$P(\vartheta)^5 = \begin{matrix} \square & & & & & \square \\ \exp(-5\vartheta) & \exp(-4\vartheta) - \exp(-5\vartheta) & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \exp(-5\vartheta) & \exp(-4\vartheta) - \exp(-5\vartheta) & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \square \exp(-5\vartheta) & \exp(-4\vartheta) - \exp(-5\vartheta) & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \square \\ \exp(-5\vartheta) & \exp(-4\vartheta) - \exp(-5\vartheta) & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \exp(-5\vartheta) & \exp(-4\vartheta) - \exp(-5\vartheta) & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \\ \exp(-5\vartheta) & \exp(-4\vartheta) - \exp(-5\vartheta) & \exp(-3\vartheta) - \exp(-4\vartheta) & \exp(-2\vartheta) - \exp(-3\vartheta) & \exp(-\vartheta) - \exp(-2\vartheta) & 1 - \exp(-\vartheta) \end{matrix}$$

las filas en $P(\vartheta)^5$ son iguales debido a que se alcanzó la distribución estacionaria, teniendo esta la forma

$$\pi(\vartheta)^T = (\exp(-5\vartheta), \exp(-4\vartheta) - \exp(-5\vartheta), \exp(-3\vartheta) - \exp(-4\vartheta), \exp(-2\vartheta) - \exp(-3\vartheta), \exp(-\vartheta) - \exp(-2\vartheta), 1 - \exp(-\vartheta))$$

Esta distribución estacionaria $\pi(\vartheta)$ también puede ser obtenida como $\pi(\vartheta)^T = e^T(I - P(\vartheta) + E)^{-1}$ (ver definición 2.1.), donde

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad E = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$e^T = \begin{matrix} \Sigma \\ 1 & 1 & 1 & 1 & 1 & 1 \end{matrix}$$

En efecto,

$$\pi^T(\vartheta) = (1 \ 1 \ 1 \ 1 \ 1 \ 1) \begin{bmatrix} \frac{2 - \exp(-\vartheta)}{1 - \exp(-\vartheta)} & 1 & 1 & 1 & 1 & \exp(-\vartheta) \\ 1 & 1 - \exp(-\vartheta) & 2 & 1 & 1 & \exp(-\vartheta) \\ 1 & 1 & 1 - \exp(-\vartheta) & 2 & 1 & \exp(-\vartheta) \\ 1 & 1 & 1 & 1 - \exp(-\vartheta) & 2 & \exp(-\vartheta) \\ 1 & 1 & 1 & 1 & 1 - \exp(-\vartheta) & \exp(-\vartheta) \\ 1 & 1 & 1 & 1 & 1 & 1 - \exp(-\vartheta) \end{bmatrix} \begin{matrix} -1 \\ \\ \\ \\ \\ 1 \end{matrix}$$

$$\pi(\vartheta)^T = (\exp(-5\vartheta), \exp(-4\vartheta) - \exp(-5\vartheta), \exp(-3\vartheta) - \exp(-4\vartheta), \exp(-2\vartheta) - \exp(-3\vartheta), \exp(-\vartheta) - \exp(-2\vartheta), 1 - \exp(-\vartheta)) .$$

Este cálculo se encuentra desarrollado en el Anexo del presente documento.

Adicionalmente, consideremos la variable aleatoria discreta $L = L_\vartheta$, cuya distribución es π , es decir, L expresa el nivel ocupado por un asegurado con frecuencia esperada de siniestros al alcanzarse la distribución estacionaria en el SBM.

3.5. Relatividades

Una relatividad, representada por la variable r_A , se define como el factor porcentual que indica cuántas veces tendrá que pagar un asegurado el valor de la prima a priori o prima base (π_{base}). En ese sentido, previo a realizar el cálculo de las relatividades para un portafolio o cartera de clientes (asegurados) según la escala del SBM definida, es necesario fijar el número de niveles A de la escala del SBM y las reglas de transición del sistema.

Para determinar las relatividades óptimas utilizaremos el enfoque de [Norberg \(1976\)](#), el cual consiste en determinar estas relatividades bajo el criterio de minimizar la exactitud predictiva del sistema, la cual viene dada por:

$$E((\Theta - r_L)^2),$$

donde Θ representa el efecto residual independiente y r_L representa la prima relativa aplicable al asegurado (después de que se haya alcanzado el estado estacionario).

Para un asegurado con parámetro estructural Θ en el portafolio de asegurados sean $\omega_k = P(\Lambda = \lambda_k)$ las probabilidades de que este asegurado sea clasificado a priori en la clase de riesgo k, la distribución de L viene dada por:

$$\begin{aligned}
P(L = A) &= \sum_{k=1}^6 P(L = A | \Lambda = \lambda_k) P(\Lambda = \lambda_k) \\
&= \sum_{k=1}^6 \omega_k \int_0^{+\infty} P(L = A | \Lambda = \lambda_k, \Theta = \theta) f(\theta) d(\theta) \\
&= \sum_{k=1}^6 \omega_k \int_0^{+\infty} \pi_A(\lambda_k \theta) f(\theta) d(\theta),
\end{aligned}$$

donde:

- $P(L = A)$ representa la probabilidad de que el asegurado se encuentre en el nivel A .
- L es el nivel ocupado por el asegurado seleccionado al azar una vez que se haya alcanzado el estado estacionario.
- k corresponde al número de clases de riesgo de la clasificación a priori del SBM propuesto.
- Λ representa la frecuencia de siniestros esperada a priori λ del asegurado.
- λ_k corresponde a la tasa de siniestros a priori de la clase de riesgo k . Se obtiene como resultado de aplicar el modelo de regresión seleccionado a los datos analizados, denominados información a priori.
- $\pi_A(\lambda_k \theta)$ corresponde a la distribución estacionaria correspondiente a A del SBM propuesto en función de λ_k y θ .
- ω_k corresponde a la probabilidad de que un asegurado sea clasificado a priori en el segmento k , cuya frecuencia de siniestros anual esperada es λ_k . También conocido como peso de la clase de riesgo k .
- $f(\theta)$ corresponde a la función de probabilidad de θ , la cual se asume como Gamma con parámetros iguales de acuerdo a la expresión indicada en 3.2, es decir, según [Denuit M. \(2007\)](#) se tiene que:

$$f(\theta) = \Gamma(a)^{(-1)} a^a \theta^{(a-1)} e^{(-a\theta)} \quad (3.6)$$

Asimismo, según [Norberg \(1976\)](#), analizando la ecuación (3.8), se puede identificar que: $E(r_L) = E(E(\Theta|L)) = E(\Theta) = 1$.

- θ corresponde a un valor de la variable aleatoria que denota el parámetro estructural y representa la tasa de siniestros, la cual considera el efecto residual **independiente asociado a los factores de riesgos no incluidos en Λ , es decir**, representa la heterogeneidad.

Adicionalmente, de acuerdo a lo indicado en [Denuit M. \(2007\)](#), consideraremos para el capítulo de aplicación que:

$$\hat{\theta} = \frac{1}{\hat{a}} \quad (3.7)$$

donde \hat{a} es la estimación del parámetro de dispersión (a) obtenido por método de máxima verosimilitud de la función binomial negativa.

Recordemos que la variable aleatoria r_L , que depende de L , más cercana a Θ en el sentido medio cuadrático viene dada por la esperanza condicional de Θ con respecto a L . Entonces nuestro objetivo es minimizar la diferencia al cuadrado esperada entre la prima relativa verdadera Θ y la prima relativa r_L aplicable al asegurado:

$$\begin{aligned} E[(\Theta - r_L)^2] &= \sum_{A=0}^{\infty} E[(\Theta - r_A)^2 | L = A] P(L = A) \\ &= \sum_{A=0}^{\infty} \int_0^{+\infty} (\Theta - r_A)^2 P(L = A | \Theta = \theta) dF_{\Theta}(\theta) \\ &= \sum_k \omega_k \int_0^{+\infty} \sum_{A=0}^{\infty} (\theta - r_A)^2 \pi_A(\lambda_k \theta) dF_{\Theta}(\theta). \end{aligned}$$

La solución está dada por la variable aleatoria $r_L = E(\Theta | L)$ la cual se caracteriza por:

$$\begin{aligned} r_L &= E(\Theta | L = A) \\ &= E(E(\Theta | L = A, \Gamma) | L = A) \\ &= \sum_k E(\Theta | L = A, \Gamma = \lambda_k) P(\Lambda = \lambda_k | L = A) \\ &= \sum_k \int_0^{+\infty} \theta \frac{P(L = A | \Theta = \theta, \Lambda = \lambda_k) \omega_k}{P(L = A, \Lambda = \lambda_k)} dF_{\Theta}(\theta) \frac{\sum P(\Lambda = \lambda_k, L = A)}{P(L = A)} \\ r_L &= \frac{\sum_k \omega_k \int_0^{\infty} \theta \pi_A(\lambda_k \theta) f(\theta) d\theta}{\sum_k \omega_k \int_0^{\infty} \pi_A(\lambda_k \theta) f(\theta) d\theta} \quad (3.8) \end{aligned}$$

Entonces, para cada nivel A en la escala del SBM, existirá una relatividad asociada r_A , la cual determinará si la prima base del conductor se ajustará como descuento o como penalización. Es decir:

- Si $r_A < 1$, el conductor recibirá un descuento sobre su prima base (π_{base}) basado en su favorable registro de siniestros pasados.
- Si $r_A = 1$, el conductor pagará lo mismo que viene pagando.
- Si $r_A > 1$, el conductor recibirá una penalización sobre su prima base (π_{base}) basado en su registro de siniestros reportados.

Finalmente, es importante resaltar la expresión simplificada para calcular la relatividad en el caso no se considere la información a priori. Ella vienen dada por:

$$r_A = \frac{\int_0^{\infty} \theta \pi_A(\lambda \theta) f(\theta) d\theta}{\int_0^{\infty} \pi_A(\lambda \theta) f(\theta) d\theta}, \quad (3.9)$$

donde λ corresponde a la tasa de siniestros esperada.



Capítulo 4

Aplicación

Con base en los modelos GLM propuestos en el capítulo 2 y al SBM propuesto en el capítulo 3, se realizará la aplicación sobre datos de seguros de automóvil con las siguientes características.

4.1. Descripción del conjunto de datos utilizado

El conjunto de datos utilizado, referenciado en [Denuit M. \(2007\)](#), consta de 155,358 registros del año 1997 de una compañía de seguros vehiculares particulares (uso personal de asegurados) en Bélgica. Cada fila o registro resume el comportamiento de un asegurado durante el periodo de un año.

A fines de conocer más el conjunto de datos, se muestran a continuación, las variables que serán de interés para esta aplicación:

| Variable | Descripción | Tipo de variable |
|----------|--|------------------|
| NCLAIMS | Número de siniestros del vehículo reportado por el asegurado durante un año | Discreta |
| AGE-PH | Edad del asegurado | Continua |
| SEX | Género del asegurado; Mujer, Hombre | Catagórica |
| SPLIT | El asegurado paga su prima en una sola cuota o lo hace en partes; En una cuota, En dos o más cuotas | Catagórica |
| SPORT | Modelo del vehículo; Deportivo, No Deportivo | Catagórica |
| FUEL | Tipo de combustible; Gasolina, Diessel | Catagórica |
| AGE-C | Antigüedad del vehículo | Discreta |
| COVERAGE | Tipo de seguro; TPL, PO-TPL+Parcial,FO-TPL+Full | Catagórica |
| FLEET | El vehículo asegurado corresponde a; Una flota, No forma parte de una flota | Catagórica |
| MONOVOL | El vehículo cuenta con ahorro de espacio; Con ahorro de espacio, Sin ahorro de espacio | Catagórica |

Cuadro 4.1: Diccionario de las variables

Para esta aplicación la variable de interés es el número de siniestros reportado por el asegurado durante el periodo de un año y el cual denotaremos como N de aquí en adelante.

Por otro lado, en el Cuadro 4.2, se aprecia el volumen de siniestros reportados por los asegurados. Como se esperaba, la cantidad de asegurados que no reportan siniestros es mucho mayor que los asegurados que reportan uno o más siniestros.

| Número de siniestros reportados | Número de asegurados con dicha condición |
|---------------------------------|--|
| 0 | 137,944 |
| 1 | 15,770 |
| 2 | 1,467 |
| 3 | 158 |
| 4 | 17 |
| 5 | 2 |

Cuadro 4.2: Número de asegurados según su número de siniestros reportados

Seguidamente, se muestran las proporciones o pesos que representan cada una de las **categorías** con respecto al total de los datos y la frecuencia anual de siniestros esperada asociada a cada **categoría** de las covariables.

| Covariable | Descripción | | Proporción % | Frecuencia siniestros % |
|------------|--|-----------------------|--------------|-------------------------|
| AGE PH | Edad del asegurado | 18-24 años | 3.5 % | 26.0 % |
| | | 25 - 30 años | 11.6 % | 19.9 % |
| | | 31 - 60 años | 63.5 % | 13.6 % |
| | | > 60 años | 21.4 % | 9.8 % |
| SEX | Género del asegurado: Mujer u Hombre | Mujer | 26.7 % | 14.8 % |
| | | Hombre | 73.3 % | 13.6 % |
| SPLIT | El asegurado paga su prima en una sola cuota o lo hace en partes 1. una cuota 2. en dos o más cuotas | Un solo pago | 49.3 % | 11.6 % |
| | | Parte el pago | 50.7 % | 16.2 % |
| SPORT | El modelo del vehículo 1. Es deportivo 2. No es deportivo | Deportivo | 9.0 % | 17.9 % |
| | | No es deportivo | 91.0 % | 13.9 % |
| FUEL | Tipo de combustible 1. Gasolina 2. Diésel | Gasolina | 69.6 % | 13.1 % |
| | | Diésel | 30.4 % | 15.7 % |
| COVERAGE | Tipo de seguro 1 TPL-Solo daño tercero, 2. PO-TPL+ Parcial daño material, 3 FO-TPL+ Full cobertura de daño material | TPL | 59.2 % | 14.6 % |
| | | PO | 28.1 % | 12.7 % |
| | | FO | 12.7 % | 13.5 % |
| AGE C | Antigüedad del vehículo | Menos de 10 años | 76.8 % | 13.1 % |
| | | Más de 10 años | 23.2 % | 15.7 % |
| FLEET | El vehículo asegurado corresponde 1 Una flota 2 No forma parte de la flota | Flota de autos | 2.8 % | 11.1 % |
| | | Sin flota | 97.2 % | 14.0 % |
| MONOVOI. | El vehículo cuenta con la parte del motor y parte delantera como un único volumen o no 0, sin ahorro de espacio 1, con ahorro de espacio | Sin ahorro de espacio | 98.9 % | 13.9 % |
| | | Con ahorro de espacio | 1.1 % | 14.2 % |

Cuadro 4.3: Análisis univariado del conjunto de datos

Del Cuadro 4.3 podemos indicar que la frecuencia anual de siniestros esperada **varía** significativamente según la variable edad, sobretodo para el grupo de 18-24 años que tiene la mayor frecuencia de siniestros, además se observa que no existe una variación significativa de la frecuencia anual de siniestros esperada en las **categorías** de las otras variables.

4.2. Clasificación a priori

Para construir la clasificación a priori, seleccionaremos primero el modelo de regresión GLM que mejor ajuste a los datos; segundo, estimaremos la frecuencia

anual de siniestros esperada luego de aplicar el modelo regresión seleccionado; y tercero, identificaremos las clases de riesgo para segmentar a priori a los asegurados según el monto de prima base a pagar. A continuación, se muestran los detalles y resultados más importantes.

4.2.1. Selección del modelo de regresión

Lo primero que se debe realizar es seleccionar el modelo de regresión que se ajuste mejor a los datos de estudio. En ese sentido, a continuación, se comparan los ajustes ofrecidos por los modelos de regresión Poisson (MRP), regresión binomial negativo (MRBN), regresión Poisson inflacionado en ceros (MRPIZ) y de regresión binomial negativo inflacionado en ceros (MRBNIZ) para los datos analizados respecto de la frecuencia anual de siniestros (FAS) observada.

| | FAS observada | FAS Esperada MRP | FAS Esperada MRBN | FAS Esperada MRPIZ | FAS Esperada MRBNIZ |
|---------|---------------|------------------|-------------------|--------------------|---------------------|
| 0 | 137,944 | 137,247.54 | 137,957.16 | 137,944.26 | 137,957.11 |
| 1 | 15,770 | 17,011.28 | 15,715.25 | 15,692.80 | 15,715.31 |
| 2 | 1,467 | 1,054.24 | 1,531.08 | 1,605.59 | 1,531.09 |
| 3 | 158 | 43.56 | 140.75 | 109.52 | 140.74 |
| 4 | 17 | 1.35 | 12.55 | 5.60 | 12.55 |
| 5 | 2 | 0.03 | 1.10 | 0.00 | 0.00 |
| valor-p | 1.00 | 0.00 | 0.20 | 0.00 | 0.00 |

Cuadro 4.4: Criterio de selección por mejor ajuste

De la tabla anterior, podemos decir que para el modelo de regresión binomial negativo (MRBN) no se rechaza la hipótesis nula, la cual afirma que los datos se ajustan a la distribución teórica. Ello considerando la prueba χ^2 y bajo un nivel de significancia de $\alpha = 0.05$. En el caso de los otros modelos de regresión como Poisson, Poisson y binomial negativo inflacionado en ceros se rechaza la hipótesis nula bajo un nivel de significancia de $\alpha = 0.05$.

En segundo lugar, para diagnosticar la presencia de sobredispersión se aplicó un test de equidispersión al modelo de regresión de Poisson de interés con ayuda de la función *dispersiontest* perteneciente a la librería AER de R, obteniéndose como resultado que se rechaza la hipótesis nula $H_0 : c = 0$ con un valor-p $< 2.2e-16$; lo que indica que la media no es igual a la varianza y además que ésta excede a la media y por tanto existe sobredispersión. Por tanto, se descarta el modelo de regresión Poisson.

En tercer lugar, analizaremos la bondad de ajuste de los modelos usando los indicadores AIC y BIC como se aprecia en la siguiente tabla:

| | AIC | BIC |
|---|-------------------|-------------------|
| Modelo de regresión Poisson | 121,650.66 | 121,660.62 |
| Modelo de regresión binomial negativo | 121,174.01 | 121,193.91 |
| Modelo de regresión binomial negativo inflacionado en ceros | 121,176.01 | 121,175.98 |
| Modelo de regresión Poisson inflacionado en ceros | 121,218.48 | 121,219.45 |

Cuadro 4.5: Resultados AIC y BIC para los modelos GLM en comparación

De la tabla anterior, se observa que el modelo de regresión binomial negativo (MRBN) presenta un mejor ajuste según el AIC y el modelo de regresión binomial

negativo inflacionado en ceros (MRBNIZ) tiene el segundo mejor BIC seguido del MRBN. En ese sentido, se confirma el descarte del modelo de regresión de Poisson y se descarta también el modelo de regresión Poisson inflacionado en ceros dado que hasta el momento han presentado resultados poco favorables en las pruebas realizadas respecto del MRBN y el MRBNIZ.

Por último, se comparará el MRBN y el MRBNIZ a través del test de Vuong:

| Vuong z-statistic | Hipótesis Alternativa | Valor-p |
|-------------------|-----------------------|----------|
| -1.924361 | MRBN > MRBNIZ | 0.027155 |

Cuadro 4.6: Resultado Test de Vuong

De la tabla anterior, se observa que bajo un nivel de significancia $\alpha = 0.05$ se rechaza la hipótesis nula, es decir, el modelo de regresión binomial negativa (MRBN) se ajusta mejor a los datos de estudio que el modelo de regresión binomial negativa inflacionado en ceros (MRBNIZ).

En conclusión, con base a todas las pruebas anteriores, se selecciona el modelo de regresión de binomial negativo (MRBN) para calcular la clasificación a priori del SBM propuesto.

4.2.2. Selección de covariables más significativas

Para seleccionar las covariables más significativas del conjunto de covariables disponibles en la base de datos analizada, se hará uso del método de *backward* para el modelo de regresión binomial negativo considerando todas las variables disponibles en la base de datos. Entonces, las variables seleccionadas para nuestro modelo son:

- Edad del asegurado (AGEPH)
- Género del asegurado (SEX)
- Modalidad de pago de prima (SPLIT)
- **Modelo deportivo del vehículo (SPORT)**
- **Tipo de combustible que usa el vehículo (FUEL)**

Considerando estas covariables seleccionadas, se procederá aplicar el modelo de regresión binomial negativo, obteniendo el siguiente resultado:

| Coefficients: | | | | | |
|----------------------|----------|------------|---------|-----------|-----|
| | Estimate | Std. Error | z value | Pr(> z) | |
| (Intercept) | -1.49069 | 0.07748 | -19.240 | <2e-16 | *** |
| EDAD 24-30 | -0.25340 | 0.03587 | -7.064 | <1.62e-12 | *** |
| EDAD 30-60 | -0.59923 | 0.03206 | -18.692 | <2e-16 | *** |
| EDAD 60-95 | -0.79567 | 0.03625 | -21.947 | <2e-16 | *** |
| SEXO Hombre | -0.03655 | 0.01692 | -2.160 | 0.0308 | * |
| SPLIT Más de 1 cuota | 0.20616 | 0.01532 | 13.444 | <2e-16 | *** |
| SPORT No | -0.17305 | 0.07150 | -2.420 | 0.0155 | * |
| FUEL Diessel | 0.14905 | 0.01615 | 9.227 | <2e-16 | *** |

Cuadro 4.7: Significancia de covariables del modelo inicial

En el cuadro anterior se observa que todas las covariables seleccionadas son significativas. En ese sentido, se observa que las covariables Sexo (Hombre) y Modelo deportivo (No) son las que tienen un valor-p mayor. Como en la **bibliografía** revisada (Antonio y Van Berkum, 2015) se aprecia que la covariable sexo es relevante para fines comerciales del SBM, entonces se mantendrá en el modelo. Sin embargo, la covariable de modelo deportivo (SPORT) se retirará del modelo más aun cuando esta no es referenciada en la **bibliografía** revisada. Considerando este cambio, realizaremos un análisis de significancia con las covariables restantes, obteniéndose:

| Coefficients: | | | | | |
|----------------------|----------|------------|---------|-----------|-----|
| | Estimate | Std. Error | z value | Pr(> z) | |
| (Intercept) | -1.65954 | 0.03394 | -48.894 | <2e-16 | *** |
| EDAD 24-30 | -0.25408 | 0.03587 | -7.083 | <1.41e-12 | *** |
| EDAD 30-60 | -0.60185 | 0.03204 | -18.786 | <2e-16 | *** |
| EDAD 60,95 | -0.79974 | 0.03621 | -22.084 | <2e-16 | *** |
| SEXO Hombre | -0.03508 | 0.01691 | -2.074 | 0.0308 | * |
| SPLIT Más de 1 cuota | 0.20604 | 0.01534 | 13.436 | <2e-16 | *** |
| FUEL Diessel | 0.14611 | 0.01610 | 9.074 | <e-16 | *** |

Cuadro 4.8: Significancia de covariables del modelo reducido

Del cuadro anterior, se observa que al retirar la covariable SPORT las covariables restantes siguen siendo significativas para modelar nuestra variable de interés.

Por otro lado, comprobaremos la calidad del ajuste del modelo inicial (modelo saturado) y el modelo reducido (modelo de interés) usando el criterio de la Devianza, el cual considera que proporción de la varianza total explica el modelo. Es decir, compararemos los resultados de la siguiente expresión:

- Modelo saturado (todas las covariables analizadas): 2 x log-likelihood: -120,126.81
- Modelo de interés (solo covariables seleccionadas): 2 x log-likelihood: -120,176.23

Del resultado anterior, se puede indicar que existe evidencia significativa de ajuste apropiado por parte de ambos modelos analizados, entonces por el criterio de

parsimonia y la Devianza elegiremos el modelo con menos covariables, es decir, el de 4 covariables.

A partir de estas covariables seleccionadas del modelo reducido, se formarán las clases de riesgo, los cuales permitirán incorporar información a priori al SBM propuesto.

4.2.3. Determinación de las clases de riesgo

Las clases de riesgo se conformarán en función al perfil de riesgo, es decir, a mayor frecuencia anual de siniestros mayor riesgo representa este grupo de personas para la empresa aseguradora. Por ello, para el presente trabajo a mayor número esperado de siniestros reportado por el asegurado más riesgosa es la clase analizada. En ese sentido, la clasificación a priori se realizará combinando todas las covariables seleccionadas en el modelo (edad del asegurado, tipo de **vehículo**, sexo, etc.), cada combinación representará una clase de riesgo (k).

Por otra parte, es importante resaltar el uso diferenciado de la covariable edad del asegurado como una variable de tipo continuo y categórico a fin de comprobar que no se pierde información o se alcanzan conclusiones diferentes. En ese sentido, en la clasificación a priori se considerarán dos escenarios, uno considerando la edad como variable continua, y otro caso considerándola como variable categórica (rango de edades). Por tanto, a continuación, se presentan las dos clasificaciones a priori que se han considerado en el presente trabajo:

- Escenario 1: clases de riesgo en base a la combinación de covariables significativas en el modelo (32 clases de riesgo), considerando la edad como variable categórica.
- Escenario 2: clases de riesgo en base a la combinación de covariables significativas en el modelo (574 clases de riesgo), considerando la edad como variable continua.

4.2.4. Determinación de prima a priori

Para el caso de las 32 clases de riesgo que se identifican a partir de la combinación de las covariables, es posible determinar la prima base, tal como se aprecia en el Cuadro 4.9 ($\pi_{base} = \lambda = E(N)$ en miles de dólares).

| N° de clases de riesgo | Edad del asegurado | Genero del asegurado | Pago de Prima | Tipo de combustible | E(N)=Prima Base | Peso en la cartera |
|------------------------|--------------------|----------------------|----------------|---------------------|-----------------|--------------------|
| clase de riesgo Nro 1 | (30, 60] | H | más de 1 cuota | gasolina | 0.1898 | 0.0044 |
| clase de riesgo Nro 2 | (60, 95] | M | 1 cuota | gasolina | 0.1705 | 0.019 |
| clase de riesgo Nro 3 | (30, 60] | H | 1 cuota | Diessel | 0.1813 | 0.074 |
| clase de riesgo Nro 4 | (60, 95] | H | 1 cuota | gasolina | 0.2099 | 0.0014 |
| clase de riesgo Nro 5 | (24, 30] | M | más de 1 cuota | gasolina | 0.1427 | 0.0231 |
| clase de riesgo Nro 6 | (24, 30] | H | más de 1 cuota | gasolina | 0.1653 | 0.1326 |
| clase de riesgo Nro 7 | (30, 60] | M | más de 1 cuota | gasolina | 0.1749 | 0.0782 |
| clase de riesgo Nro 8 | (60, 95] | H | más de 1 cuota | gasolina | 0.2022 | 0.1537 |
| clase de riesgo Nro 9 | (24, 30] | M | 1 cuota | gasolina | 0.1041 | 0.0945 |
| clase de riesgo Nro 10 | (30, 60] | H | 1 cuota | gasolina | 0.1208 | 0.0206 |
| clase de riesgo Nro 11 | (30, 60] | M | 1 cuota | gasolina | 0.1282 | 0.0024 |
| clase de riesgo Nro 12 | (16, 24] | M | 1 cuota | gasolina | 0.2194 | 0.0048 |
| clase de riesgo Nro 13 | (30, 60] | M | más de 1 cuota | Diessel | 0.1481 | 0.0048 |
| clase de riesgo Nro 14 | (30, 60] | H | más de 1 cuota | Diessel | 0.1006 | 0.0019 |
| clase de riesgo Nro 15 | (16, 24] | H | más de 1 cuota | gasolina | 0.1166 | 0.0975 |
| clase de riesgo Nro 16 | (60, 95] | M | más de 1 cuota | gasolina | 0.1236 | 0.0057 |
| clase de riesgo Nro 17 | (24, 30] | M | más de 1 cuota | Diessel | 0.1429 | 0.021 |
| clase de riesgo Nro 18 | (24, 30] | H | 1 cuota | Diessel | 0.085 | 0.0447 |
| clase de riesgo Nro 19 | (30, 60] | M | 1 cuota | Diessel | 0.0988 | 0.0118 |
| clase de riesgo Nro 20 | (16, 24] | H | 1 cuota | gasolina | 0.1051 | 0.0015 |
| clase de riesgo Nro 21 | (24, 30] | H | 1 cuota | gasolina | 0.1215 | 0.0058 |
| clase de riesgo Nro 22 | (60, 95] | M | 1 cuota | Diessel | 0.0825 | 0.0032 |
| clase de riesgo Nro 23 | (60, 95] | H | más de 1 cuota | Diessel | 0.2344 | 0.0176 |
| clase de riesgo Nro 24 | (60, 95] | H | 1 cuota | Diessel | 0.0958 | 0.0093 |
| clase de riesgo Nro 25 | (24, 30] | H | más de 1 cuota | Diessel | 0.1016 | 0.0042 |
| clase de riesgo Nro 26 | (16, 24] | H | más de 1 cuota | Diessel | 0.1175 | 0.0127 |
| clase de riesgo Nro 27 | (16, 24] | M | más de 1 cuota | gasolina | 0.2696 | 0.0057 |
| clase de riesgo Nro 28 | (16, 24] | M | 1 cuota | Diessel | 0.1837 | 0.0177 |
| clase de riesgo Nro 29 | (16, 24] | H | 1 cuota | Diessel | 0.2123 | 0.0117 |
| clase de riesgo Nro 30 | (24, 30] | M | 1 cuota | Diessel | 0.2258 | 0.0292 |
| clase de riesgo Nro 31 | (60, 95] | M | más de 1 cuota | Diessel | 0.2612 | 0.0174 |
| clase de riesgo Nro 32 | (16, 24] | M | más de 1 cuota | Diessel | 0.1476 | 0.0604 |

Cuadro 4.9: Clase de riesgo - Escenario 1

Para el caso de las 574 clases de riesgo que se identifican a partir de la combinación de la variable edad (variable continua) con el resto de covariables. Ver Cuadro 4.10.

| N° de clases de riesgo | Edad del asegurado | Genero del asegurado | Pago de Prima | Tipo de combustible | E(N)=Prima Base | Peso en la cartera |
|-------------------------|--------------------|----------------------|----------------|---------------------|-----------------|--------------------|
| clase de riesgo Nro 1 | 50 | H | más de 1 cuota | gasolina | 0.121 | 0.0062 |
| clase de riesgo Nro 2 | 64 | M | 1 cuota | gasolina | 0.1483 | 0.0059 |
| clase de riesgo Nro 3 | 60 | H | 1 cuota | Diessel | 0.1539 | 0.5353 |
| : | : | : | : | : | : | : |
| clase de riesgo Nro 555 | 88 | M | más de 1 cuota | gasolina | 0.1191 | 0.0037 |
| clase de riesgo Nro 556 | 92 | M | 1 cuota | gasolina | 0.1265 | 0.0037 |
| clase de riesgo Nro 557 | 79 | H | más de 1 cuota | Diessel | 0.0903 | 0.0037 |
| clase de riesgo Nro 558 | 19 | M | 1 cuota | Diessel | 0.073 | 0.0037 |
| clase de riesgo Nro 559 | 86 | M | más de 1 cuota | gasolina | 0.093 | 0.0037 |
| clase de riesgo Nro 560 | 89 | H | 1 cuota | Diessel | 0.0741 | 0.0036 |
| clase de riesgo Nro 561 | 82 | M | más de 1 cuota | gasolina | 0.0955 | 0.0036 |
| clase de riesgo Nro 562 | 18 | H | 1 cuota | gasolina | 0.1745 | 0.0036 |
| clase de riesgo Nro 563 | 18 | M | más de 1 cuota | gasolina | 0.1394 | 0.0035 |
| clase de riesgo Nro 564 | 90 | M | más de 1 cuota | gasolina | 0.0944 | 0.0059 |
| clase de riesgo Nro 565 | 80 | M | 1 cuota | Diessel | 0.0945 | 0.0035 |
| clase de riesgo Nro 566 | 18 | H | 1 cuota | Diessel | 0.1294 | 0.0035 |
| clase de riesgo Nro 567 | 93 | M | 1 cuota | gasolina | 0.08 | 0.0035 |
| clase de riesgo Nro 568 | 85 | H | más de 1 cuota | Diessel | 0.0724 | 0.0035 |
| clase de riesgo Nro 569 | 93 | H | 1 cuota | Diessel | 0.1376 | 0.0035 |
| clase de riesgo Nro 570 | 84 | H | 1 cuota | Diessel | 0.1274 | 0.0034 |
| clase de riesgo Nro 571 | 79 | M | 1 cuota | Diessel | 0.1633 | 0.0034 |
| clase de riesgo Nro 572 | 92 | H | más de 1 cuota | gasolina | 0.156 | 0.0034 |
| clase de riesgo Nro 573 | 17 | M | más de 1 cuota | Diessel | 0.1345 | 0.0034 |
| clase de riesgo Nro 574 | 88 | H | más de 1 cuota | Diessel | 0.1932 | 0.0034 |

Cuadro 4.10: Clase de riesgo - Escenario 2

Cabe resaltar que estas clases de riesgo se han identificado a partir de un análisis univariado, selección de variables y consideraciones indicadas en la principal bibliografía consultada (Antonio y Van Berkum, 2015). Asimismo, la determinación de

estas clases de riesgo tiene la ventaja de ser bastante intuitiva para el área de ventas de la aseguradora e incluso le permite clasificar claramente a un nuevo cliente (asegurado) del cual no se cuenta con historial previo de siniestros reportados.

4.3. Clasificación a posteriori

El SBM propuesto hace uso de la clasificación a priori, específicamente incorpora la frecuencia anual de siniestros esperado (λ) y pesos obtenidos (ω), a fin de incorporar la información a posteriori y obtener la prima ajustada ($\pi_{ajustada}$).

De acuerdo a la ecuación (3.8), y para 32 clases de riesgo a priori, la fórmula que se usó para el cálculo de la relatividad considerando la información a priori es:

$$r_A = \frac{\sum_{k=1}^{32} \omega_k \int_0^{\infty} \theta \pi_A(\lambda_k \theta) f(\theta) d\theta}{\sum_{k=1}^{32} \omega^k \int_0^{\infty} \pi^A(\lambda^k \theta) f(\theta) d\theta},$$

donde:

- k corresponde a las 32 clases de riesgo analizadas.
- ω es el vector de pesos, dado por

$$\omega = (0.0044, 0.019, 0.074, 0.0014, 0.0231, 0.1326, 0.0782, 0.1537, 0.0945, 0.0206, 0.0024, 0.0048, 0.0000, 0.0019, 0.0975, 0.0057, 0.021, 0.0447, 0.0118, 0.0015, 0.0058, 0.0032, 0.0176, 0.0093, 0.0042, 0.0127, 0.0057, 0.0177, 0.0117, 0.0292, 0.0174, 0.0604)$$
- λ es el vector de frecuencia anual de siniestros esperado, dado por

$$\lambda = (0.1898, 0.1705, 0.1813, 0.2099, 0.1427, 0.1653, 0.1749, 0.2022, 0.1041, 0.1208, 0.1282, 0.2194, 0.1481, 0.1006, 0.1166, 0.1236, 0.1429, 0.085, 0.0988, 0.1051, 0.1215, 0.0825, 0.2344, 0.0958, 0.1016, 0.1175, 0.2696, 0.1837, 0.2123, 0.2258, 0.2612, 0.1476)$$
- θ , que corresponde al valor del parámetro de heterogeneidad, el cual (para los datos analizados en el presente trabajo) se estimará luego de aplicar la ecuación (3.7), es decir: $\hat{\theta} = 1,667 = \frac{1}{\hat{a}} = \frac{1}{0.599}$
- $f(\theta)$ corresponde a la función de θ señalada en (3.6), la cual se encuentra en el código R como la variable *f.dist* en el Anexo del presente documento.
- $\pi_A(\lambda_k \theta)$ corresponde a la distribución estacionaria correspondiente a A del SBM propuesto en función de λ_k y θ , donde A es el estado o nivel: 0, 1, 2, 3, 4, y 5 del SBM propuesto. Asimismo, para mayor detalle de cómo se calcula la distribución estacionaria en el presente trabajo, ver función *lim.distr* en el Anexo - Código R.

De similar manera, se ha procedido en el presente trabajo para ω y λ en el modelo con 574 clases de riesgo a priori.

Asimismo, es importante precisar que el SBM propuesto tiene la capacidad de operar de dos maneras: incorporando información a priori, y sin incorporar información a priori. En ese sentido, y de acuerdo a la ecuación (3.9), la fórmula que se usó para el cálculo de la relatividad sin considerar la información a priori es:

$$r_A = \frac{\int_0^{\infty} \theta \pi_A(\lambda\theta) f(\theta) d\theta}{\int_0^{\infty} \pi_A(\lambda\theta) f(\theta) d\theta}$$

donde en la fórmula anterior del cálculo de la relatividad se tiene que: λ es la tasa de siniestros esperada (estimada como el promedio de las tasas de siniestros).

Por tanto, y en base al modelo propuesto en el capítulo anterior, a continuación, se muestran los resultados del SBM considerando los valores de las relatividades (r_A) tanto para cuando se incorpora información a priori como cuando no se incorpora dicha información para cada nivel de “ $-1/topscale$ ” (Ver el Cuadro 4.11).

| Nivel en “-1/topscale” | Relatividad | | | |
|---------------------------|--------------------------------------|--|--------------------------------------|--|
| | Edad como variable continua | | Edad como variable categórica | |
| | considerando información a priori | sin considerar información a priori | considerando información a priori | sin considerar información a priori |
| 0 | 47.351 % | 49.193 % | 44.721 % | 49.190 % |
| 1 | 130.237 % | 138.554 % | 123.828 % | 138.545 % |
| 2 | 146.484 % | 155.264 % | 139.914 % | 155.255 % |
| 3 | 167.59 % | 176.574 % | 161.24 % | 176.566 % |
| 4 | 196.229 % | 204.701 % | 191.181 % | 204.694 % |
| 5 | 237.564 % | 243.554 % | 237.229 % | 243.552 % |

Cuadro 4.11: Resumen de resultados de las relatividades obtenidas por el SBM propuesto

Estas relatividades permitirán calcular la prima ajustada para cada asegurado según el nivel en que se encuentre actualmente. Como se aprecia en la tabla cuando no se incorpora información a priori los ajustes realizados a la prima base son mayores y con tendencia al alza que cuando sí se incorpora información a priori. En ese sentido, se aprecia que el modelo incorporando información a priori fortalecerá la retención de clientes dado que los precios de las primas no serán excesivamente altos y premiará a los mejores clientes (asegurados que NO presentan siniestros) y castigará a los malos clientes (asegurados que presentan siniestros).

Finalmente, para fines de mostrar de forma referencial cómo se visualizaría la tarifación en el SBM propuesto, a continuación, se muestra para 5 asegurados de ejemplo, su prima base, el nivel inicial en el SBM, el número de reclamaciones en el año analizado, el nivel actual en el SBM en función al número de reclamaciones en el año, la relatividad asociada al nivel SBM actualizado y la prima ajustada ($\pi_{ajustada}$) en base al SBM establecido. Ver Cuadro 4.12.

| Asegurado | Prima base | Nivel inicial en SBM | Número de reclamaciones en el año | Nivel actualizado en SBM | Relatividad asociado al Nivel del SBM | Prima ajustada (Prima base* Relatividad) |
|-------------|------------|----------------------|-----------------------------------|--------------------------|---------------------------------------|--|
| Asegurado 1 | 120.00 USD | 1 | 0 | 0 | 63 % | 75.60 USD |
| Asegurado 2 | 120.00 USD | 1 | 1 | 5 | 186 % | 223.20 USD |
| Asegurado 3 | 125.00 USD | 5 | 0 | 4 | 166 % | 207.50 USD |
| Asegurado 4 | 150.00 USD | 2 | 0 | 1 | 128 % | 192.00 USD |
| Asegurado 5 | 175.00 USD | 1 | 0 | 0 | 63 % | 110.25 USD |

Cuadro 4.12: Aplicación práctica con el modelo propuesto

Nota: Apreciar los dos primeros casos, si bien ambos tienen la misma prima base asignada al inicio del periodo analizado, luego de un periodo por los siniestros reportados, el primer caso es premiado disminuyendo de nivel en el SBM y pagando el 63 % del monto de la prima base original, en cambio el segundo caso es castigado aumentando el nivel en el SBM y pagando 186 % del valor de la prima base asignado inicialmente.



Capítulo 5

Conclusiones

5.1. Conclusiones

- En este trabajo se desarrolló un SBM con información a priori y a posteriori, mostrándose cómo se utilizan ambas informaciones por separado y en conjunto en la determinación de primas más justas para los asegurados de un producto de seguros vehiculares. Ello, dado que, el asegurado que presente un comportamiento sin siniestros pagará menos que un asegurado que presente siniestros en el periodo evaluado.
- El SBM desarrollado en la aplicación de este trabajo que considera información a priori mostró relatividades más competitivas que el SBM que no considera información a priori. Esta mayor precisión en el cálculo de la prima ajustada del asegurado contribuye en no elevar el indicador de fuga de clientes por precios poco competitivos (primas muy por encima del promedio del mercado de manera no necesariamente justificada).
- Para el modelamiento de datos de conteo, como es el caso de número de siniestros reportados por una cartera de asegurados, es importante verificar la presencia de sobredispersión a fin de asegurar que el modelo seleccionado soporte esta característica en los datos analizados. Asimismo, como se apreció en la sección de aplicación no siempre los modelos inflacionados en ceros son mejor que sus versiones originales, tal es el caso del presente trabajo donde el modelo de regresión binomial negativo resultó ganador respecto de su versión inflacionada en ceros.
- Para la comparación de modelos es bastante útil el método de Vuong dado que permite tener una opción alternativa al AIC y BIC, sobre todo cuando estos anteriores no muestran un claro modelo ganador.
- Para la selección de variables, es importante realizar un análisis de datos (univariado / multivariado) con el objetivo de visualizar el poder de discriminación que puede tener una variable respecto de la frecuencia de siniestros de los asegurados. Por ejemplo, como se aprecia en los resultados, la variable edad influye significativamente en la variable de respuesta que deseamos modelar.

Asimismo, es relevante realizar análisis de significancia con el objetivo de solo quedarse con las variables más relevantes para el modelo.

- En la sección aplicación cuando no se considera información a priori, se aprecia que las relatividades de la regresión binomial negativa con variable edad continua respecto a esta misma regresión con variable edad categórica presenta diferencias **mínimas** en términos solo de decimales.
- En la sección aplicación cuando se considera información a priori, se aprecia que las relatividades de la regresión binomial negativa con variable edad continua respecto a esta misma regresión con variable edad categórica son mayores. **Específicamente**, RBN con edad categórica premia a las personas que no reportan siniestros y castiga al que reporta siniestros de manera más severa e incremental conforme se aumentan los niveles en el SBM.
- Un SBM puede contar con el número de niveles que la empresa desee, solo dependerá de definir las reglas de transición y sus respectivas probabilidades de transición.

5.2. Recomendaciones

- Extender el presente SBM a través del uso de modelos lineales generalizados mixtos (GLMM) para el tratamiento de datos longitudinales. En el presente trabajo se utilizó datos transversales.
- Adicionar a las distribuciones estudiadas de conteo (Poisson, binomial negativo, distribuciones inflacionados en ceros), otros modelos alternativos existentes o emergentes.
- Adicionar otras formas de determinar los grupos de riesgos a priori, por ejemplo, a través de la teoría del análisis de conglomerados o clusters. Asimismo, evaluar el uso de cuantiles o técnica equivalente.
- Extender el sistema estudiado considerando covariables concretas que permitan predecir el monto de siniestro (reportado por el asegurado) y no solo la frecuencia.
- Calibrar el modelo propuesto con datos de seguros vehiculares de Perú.

Apéndice A

Anexo - Código

A.1. Anexo - Código R

```
#Lectura de base de datos de asegurados
Datos_Vehiculos <- read.table("MTPL_Data_1997.txt",
                             sep="\t", dec=".", header = TRUE)

# Modelo inicial
fit1=glm(NCLAIMS~AGEPH+AGEC+FUEL+PERIOD+COVERAGE+SPORT+FLEET+MONOVOL+
        SEX,family=poisson,data=Datos1)
summary(fit1)
exp(coef(fit1))
###Selección de Variables
library(MASS)
stepAIC(fit1)

#Modelo con variables seleccionadas
fit2=glm(NCLAIMS~AGEPH+FUEL+PERIOD+COVERAGE,
        family=poisson,data=Datos1)
summary(fit2)
library(AER)
dispersiontest(fit2,trafo=1)

fit3=zeroinfl(NCLAIMS~AGEPH+FUEL+PERIOD+COVERAGE+SPORT,data=Datos1,
             dist="poisson")
summary(fit3)
fit4=zeroinfl(NCLAIMS~AGEPH+PERIOD+COVERAGE+FUEL+COVERAGE+SEX,
             data=Datos1,dist="poisson")
summary(fit4)
vuong(fit2,fit4)

help(zeroinfl)

fit5=zeroinfl(NCLAIMS~AGEPH+PERIOD+COVERAGE|1,
             data=Datos1,dist="negbin")
```

```

summary(fit5)
fit6=glm.nb(NCLAIMS~AGEPH+FUEL+PERIOD+COVERAGE,data=Datos1)

# Se obtiene como mejor modelo a la binomial negativa
vuong(fit2, fit6)
vuong(fit4, fit6)
vuong(fit5, fit6)

# Incorporando información a priori al SBM propuesto

# Lambda media
Datos3$lambda_medio <- Datos3 %>% group_by(clase) %>% summarise(Lambda_
  medio=mean(Lambda))

# lambda medio por cada cluster le calcula la media de su lamda , %>%
  embudo
## esta es!!!
tabla_1 <- Datos3[,list(weights=min(W),lambda=mean(Lambda)),by=c("clase")
  ] %>% setorder(clase)
weights=c(tabla_1$weights) #lo guarde como vector
lambda=c(tabla_1$lambda) # promedio de los lamdas por categorías, lo guardo
  como vector

write.csv(weights,"weights_clases32.csv")
write.csv(lambda,"lambda_clases32.csv")

n=length(weights)
n=length(lambda)

# Código R correspondiente a la clasificación posteriori con el SBM
  propuesto
# aqui se construye la matriz de 1 paso P(theta)
Pmatrix =
  function(th) {
    P = matrix(nrow=6, ncol=6, data=0)
    P[1,1]=P[2,1]=P[3,2]=P[4,3]=P[5,4]=P[6,5]= exp(-th)
    P[,6] = 1-exp(-th)
    return(P)}

# aqui se consigue la distribución estacionaria pi(theta)
lim.distr =
  function(matrix) {
    et = matrix(nrow=1, ncol=dim(matrix)[2], data=1)
    E = matrix(nrow=dim(matrix)[1], ncol=dim(matrix)[2], data=1)
    mat = diag(dim(matrix)[1]) - matrix + E
    inverse.mat = solve(mat)
    p = et %*% inverse.mat
  }

```

```

return(p)}

## relatividad con info apriori.
a=1/1.6668
n=length(weights)

int3 =
function(theta, lambda, a, l) {
  p = lim.distr(Pmatrix(lambda*theta))
  f.dist = gamma(a)^(-1) * a^a * theta^(a-1) * exp(-a*theta)
  return(theta*p[1,l+1]*f.dist)}
int4 =
function(theta, lambda, a, l) {
  p = lim.distr(Pmatrix(lambda*theta))
  f.dist = gamma(a)^(-1) * a^a * theta^(a-1) * exp(-a*theta)
  return(p[1,l+1]*f.dist)}
teller1 = teller2 = noemer = array(dim=6, data=0)
result1 = result2 = array(dim=6, data=0)

### With a priori ratemaking !!!
for (i in 0:5) {
  b = c = array(dim=n,data=0)
  for (j in 1:n) {
    b[j] = integrate(Vectorize(int3),lower=0, upper=Inf,lambda=lambda[j],a=a
, l=i)$value
    c[j] = integrate(Vectorize(int4),lower=0, upper=Inf,lambda=lambda[j],a=a
, l=i)$value
  }
  teller1[i+1] = b %*% weights
  noemer[i+1] = c %*% weights
  R = teller1/noemer
}
R
class(R)
R_apriori <- data.table(l=c(0:5),rl=R*100)

##### Sin a priori ratemaking
a.hat = 1/1.6668
lambda.hat = mean(Datos3$Lambda) #promedio de los lambdas generados para
cada obervacion con MRBN
#este es el lamda estimado de toda la data osea, un promedio

int1 =
function(theta, s, a, lambda) {
  a = a.hat
  lambda = lambda.hat
  f.dist = gamma(a)^(-1) * a^a * theta^(a-1) * exp(-a*theta)

```

```

p = lim.distr(Pmatrix((lambda*theta)))
return(theta*p[1,s+1]*f.dist)}
P1 = matrix(nrow=1, ncol=6, data=0)
for (i in 0:5) P1[1,i+1] = integrate(Vectorize(int1),lower=0,upper=Inf,s=i)
$value

int2 =
function(theta, s, a, lambda) {
  a = a.hat
  lambda = lambda.hat
  f.dist = gamma(a)^(-1) * a^a * theta^(a-1) * exp(-a*theta)
  p = lim.distr(Pmatrix((lambda*theta)))
  return(p[1,s+1]*f.dist)}
P2 = matrix(nrow=1, ncol=6, data=0)

for (i in 0:5) P2[1,i+1] = integrate(Vectorize(int2),lower=0,upper=Inf,s=i)
$value
R_SinApriori = P1 / P2
R_SinApriori

R_SinApriori <- c(R_SinApriori) #convierte de matriz a vector
data.table(l=c(0:5),rl=R_SinApriori*100)

```

A.2. Anexo - Código MATLAB

```

% Sean los vectores:

% e: vector de 1's
% I: matriz identidad
% P: matriz de probabilidades de transición
% E: matriz de 1's

e =

    1  1  1  1  1  1

I =

    1  0  0  0  0  0
    0  1  0  0  0  0
    0  0  1  0  0  0
    0  0  0  1  0  0
    0  0  0  0  1  0
    0  0  0  0  0  1

```

```
P =
[ exp(-th), 0, 0, 0, 0, 1 - exp(-th)]
[ exp(-th), 0, 0, 0, 0, 1 - exp(-th)]
[ 0, exp(-th), 0, 0, 0, 1 - exp(-th)]
[ 0, 0, exp(-th), 0, 0, 1 - exp(-th)]
[ 0, 0, 0, exp(-th), 0, 1 - exp(-th)]
[ 0, 0, 0, 0, exp(-th), 1 - exp(-th)]
```

```
E =
1 1 1 1 1 1
1 1 1 1 1 1
1 1 1 1 1 1
1 1 1 1 1 1
1 1 1 1 1 1
1 1 1 1 1 1
```

% La distribución estacionaria pi puede ser obtenida como $\pi = e \cdot (I - P + E)^{-1}$
 % a continuación tenemos que $(I - P + E)$ es igual a:

```
(I-P+E) = simplify(I-P+E)
[ 2 - exp(-th), 1, 1, 1, 1, exp(-th)]
[ 1 - exp(-th), 2, 1, 1, 1, exp(-th)]
[ 1, 1 - exp(-th), 2, 1, 1, exp(-th)]
[ 1, 1, 1 - exp(-th), 2, 1, exp(-th)]
[ 1, 1, 1, 1 - exp(-th), 2, exp(-th)]
[ 1, 1, 1, 1, 1 - exp(-th), 1 + exp(-th)]
```

% para obtener la distribución estacionaria pi debemos multiplicar el
 vector e por la
 % inversa de la matriz $(I - P + E)$ y el resultado es el siguiente: pi

```
= simplify(e*(I-P+E)^(-1)) =
[ exp(-5*th), exp(-5*th)*(exp(th) - 1), exp(-4*th)*(exp(th) - 1), exp(-3*th)
)* (exp(th) - 1), exp(-th) - exp(-2*th), 1 - exp(-th)]
```

Bibliografía

- Alcaide, M. (2015). *Modelo de regresión binomial negativa*, Master's thesis, Universidad de Sevilla. Tesis de matemáticas.
- Andrea Camargo, L. G. F. G. (2014). Fomento del acceso a los seguros en el Perú. https://a2ii.org/sites/default/files/reports/fomento_del_acceso_a_los_seguros_en_el_peru.pdf.
- Antonio, K, S. P. y Van Berkum, F. (2015). *Longitudinal Data and Experience Rating in Computational Actuarial Science with R*, CRS Press.
- Bozdogan, H. (2000). Akaike information criterion and recent developments in information complexity, *Journal of Mathematical Psychology* 44(1): 62–91.
- Cameron, A. y Trivedi, P. (1998). *Regression Analysis of Count Data*, Cambridge University Press.
- Cordeiro G.M., V. K. y. B. L. (2000). Improved estimation for robust econometric regression models, *Brazilian Journal of Probability and Statistics* 14: 141–157.
- Denuit M., Marchal X, P. S. W. J. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*, Wiley.
- Gevorg, G. A. (2016). Martingale approach to bonus-malus systems. Disertacion.
- Haggstrom (2002). *Finite Markov Chains and Algorithmic Applications*, Cambridge University Press.
- Lambert, D. (1992). Zero-inflated poisson regression with an application to defects in manufacturing, *Technometrics* 34: 1–14.
- Lemaire, J. (1995). *Bonus-Malus Systems in Automobile Insurance*, Springer Science+Business Media, LLC.
- Monzon, A. G. (2009). *Inference and diagnosis in models for counting data with excess of zeros*, Master's thesis, Universidade Estadual de Campinas. Tesis de maestría.
- Norberg, R. (1976). Acredibility theory for automobile bonus systems, *Scandinavian Actuarial Journal* 1: 92–107.
- Paula, G. A. (2013). *Modelos de regresión con apoyo computacional*, Universidad de Sao Paulo.
- Pinquet, J. (1997). Allowance for cost of claims in bonus-malus systems, *ASTIN BULLETIN* 27: 38–41.
- Pinto, F. A. S. (2014). *Simulación en sistemas de bonus malus*, Master's thesis, Universidad Nova de Lisboa. Disertación.

Tzougas, G., V. S. y Frangos, N. (2007). Optimal bonus-malus systems using generalized additive models for location, scale and shape, *Statistics* 1: 1–17.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *The Econometric Society (Econometrica)* 57(2): 307–333.

