UNIVERSIDAD AUTÓNOMA METROPOLITANA
UNIDAD IZTAPALAPA
CASA ABIERTA AL TIEMPO

Análisis de las Primas de Riesgo en Seguros de Automóviles: Una Aplicación de los Modelos Lineales Generalizados

Tesis que presenta:
Adrián Martínez Gutiérrez

Para obtener el Grado de:
Maestro en Ciencias
Matemáticas Aplicadas e Industriales

Asesora de Tesis:
Dra. Blanca Rosa Pérez Salvador
Co-Asesor de Tesis:
M. en C. Alejandro Román Vásquez

Jurado Calificador:

Presidente: Dr. Francisco Ariza Hernández

Secretario: Dr. Alberto Castillo Morales

Vocal: Dra. Blanca Rosa Pérez Salvador

Vocal: Mtro. Carlos Omar Jiménez Palacios

Vocal: M. en C. Alejandro Román Vásquez

Ciudad de México, Abril 2017

Agradecimientos

La presente tesis fue un esfuerzo de varias personas que estuvieron opinando, dándome ánimo y apoyándome para concluir este trabajo

A mis papás Irma y Filiberto por todo su apoyo incondicional

A mis hermanos Margarita y Alberto por la paciencia que me han tenido

A mis compañeros y amigos de la MCMAI por todos los momentos compartidos

A mis amigos de INEGI que me apoyaron en la última parte de este proceso

A mis asesores:
Dra. Blanca Rosa Pérez Salvador y al M. en C. Alejandro Román Vázquez
por el tiempo dedicado en la dirección de este trabajo

A los miembros del Jurado:
Dr. Francisco Ariza Hernández, Dr. Alberto Castillo Morales
y al Mtro. Carlos Omar Jiménez Palacios por las opiniones y sugerencias
para que este trabajo terminara de mejor manera

A todos muchas gracias

Adrián

Índice general

Introducción

A raíz de la crisis financiera que azotó el mundo a partir del año 2008, el sector financiero reforzó las medidas necesarias que se venían desarrollando para cuantificar de una manera más eficiente y segura los posibles riesgos a los que están sujetos. En este sentido, el sector asegurador, dado su espíritu intrínsicamente económico, no se quedó atrás y buscó implementar acciones para contener los riesgos a los que las compañías aseguradoras están expuestas. A partir del 2009, comenzó a gestarse en Europa una serie de nuevas normas directivas basadas principalmente en tres aspectos: una mejor medición de activos, pasivos y capital; un proceso de supervisión más eficiente; y finalmente, requerimientos de transparencia de la información. Estos cambios, fueron retomados por la industria mexicana de seguros, los cuales han sido plasmados en la nueva Ley de Instituciones de Seguros y de Fianzas, la cuál empezó a operar a partir del 4 de abril del 2015.

En este contexto de cambios en el sector asegurador es fundamental una medición más rigurosa y detallada de los riesgos a los que están sujetas las compañías para reducir las probabilidades de insolvencia de las aseguradoras. Dentro de los requerimientos cuantitativos fundamentales, está la formulación de esquemas más exactos para el cálculo de las reservas técnicas, las cuales tienen que ver con el capital que las compañías deben reservar para afrontar sus obligaciones, entre las que destacan el pago de siniestros a los asegurados por los diversos riesgos contratados en una póliza de seguros. Por tal motivo, las aseguradoras deben tener técnicas estadísticas y actuariales que les permitan estimar de manera muy precisa el monto de capital que van a requerir para afrontar estos compromisos contractuales, lo que se traduce en la formulación de primas de riesgo más exactas. En este marco, queda de manifiesto la importancia de generar modelos que permitan estimar mejor las primas de riesgo de las compañías aseguradoras.

En la actualidad, para el cálculo de la prima de riesgo, uno de los métodos estadísticos más comúnmente usado es el producto de las frecuencias (número de siniestros entre unidades expuestas) por la severidad (monto de los siniestros entre el número de siniestros). Cuando se tienen grupos o cortes de información con la suficiente robustez estadística, éste método es adecuado para obtener estimaciones aceptables; sin embargo, cuando se desea aproximar una prima de riesgo a un nivel más granular, este procedimiento puede arrojar cálculos con una variabilidad muy

grande, debido al reducido número de observaciones que puede existir en ciertos cortes de información. Es aquí donde los Modelos Lineales Generalizados (GLM's por sus siglas en inglés) juegan un papel importante, pues con ellos se pueden tener mejores estimaciones para evitar dicha volatilidad.

Los GLM's son básicamente regresiones lineales, donde existe una variable respuesta (también conocida como variable dependiente), de la cual se busca describir su comportamiento a través de variables explicativas (también conocidas como variables independientes). Cuando la variable respuesta posee un comportamiento que se puede modelar empleando una distribución normal, se usa la regresión lineal convencional. Sin embargo, si la variable dependiente tiene una distribución diferente es mejor usar la generalización de los modelos lineales, que incluye otros tipos de distribuciones que pertenecen a la familia exponencial. Para el caso que compete este proyecto, las variables respuesta como la frecuencia y la severidad de los siniestros, típicamente no siguen una distribución normal, pero si alguna de la familia exponencial. Dentro de todo este contexto, para tener estimaciones con mayor robustez y significancia, se propone analizar la información de todo el sector asegurador.

Un punto importante que se debe tener en consideración para la buena estimación de las primas de riesgo es la antiselección o selección adversa, que consiste en cobrar una misma tarifa a los asegurados con diferente exposición al riesgo, es por esto la importancia de utilizar técnicas estadísticas para agrupar los riesgos con pérdidas potenciales similares para evitar esta situación.

El problema que se propone resolver y objetivo de esta tesis es generar un modelo para describir la prima de riesgo de las principales coberturas (Daños Materiales, Robo Total, Responsabilidad Civil por Daños a Terceros y Gastos Médicos a Ocupantes) que son contratadas por los propietarios de vehículos residentes de la República Mexicana empleando la información del sector asegurador de tal manera que se evite la antiselección o selección adversa ya que con esto también se estará evitando la volatilidad que presentan los métodos más comunes para dicha estimación. De esta manera, los resultados obtenidos con este trabajo serán trascendentales para todo el sector, y pueden ser usados como puntos de referencia, calibración y estimación de las primas de riesgo de cada compañía de la industria aseguradora mexicana.

La importancia de generar este modelo radica en que la buena estimación de la prima pura de riesgo permitiría obtener tarifas suficientes las cuales incrementarían la rentabilidad de la aseguradora y así tener la solvencia necesaria para poder afrontar los costos generados por los siniestros, por el contrario, una mala modelación impactaría en la tarifa comercial, lo que probablemente se tendría una tarifa insuficiente la cual implicaría problemas técnicos como menos rentabilidad o las utilidades resultantes no sean las esperadas inclusive se pondrían presentar pérdidas, por otra parte también repercutiría en problemas comerciales ya que se podría exceder el costo de la tarifa y por lo cual perdería competitividad en el sector asegurador.

Este trabajo está conformado de tres capítulos. En el primer capítulo se pre-

sentan los conceptos teóricos para formular los modelos de las primas de riesgo. En el segundo capítulo se presentan los modelos y el proceso de tarificación, en el tercer capítulo se presenta una aplicación con datos reales del cálculo de las primas de riesgo para las cuatro principales coberturas (Daños Materiales, Robo Total, Responsabilidad Civil y Gastos Médicos) de una póliza de seguro de automóvil; Finalmente, se exponen las conclusiones a las que se llegan según el objetivo del presente proyecto; en los anexos se presentan las salidas de los modelos obtenidas del software estadístico R [15].

Capítulo 1
Preliminares

En este capítulo se hará una breve descripción de algunos métodos estadísticos multivariantes que nos ayudarán más adelante en la forma de cuantificar la prima de riesgo en seguros de automóviles.

Estos métodos son de suma importancia, ya que se debe contar con una forma clara y eficaz para trabajar con datos multivariados. Los temas que se trataran en una sección son:

- **La familia exponencial:** Debido a que se trabajará con distintas distribuciones de probabilidad, podemos considerar a una única familia paramétrica de funciones que contiene todas las propiedades que nos interesa.
- **Modelos lineales generalizados:** En este caso, se tiene la necesidad de hacer regresiones lineales con variables aleatorias que no se distribuyen de manera normal.
- **Análisis de conglomerados:** Debido a que se trabajará con una población de elementos con características diferentes, y se tendrá la necesidad de formar grupos lo más homogéneos posible.

1.1. La Familia Exponencial

Sea Y una variable aleatoria cuya distribución de probabilidad depende de un solo parámetro θ. La distribución de la variable Y pertenece a la familia exponencial si puede expresarse de la forma

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \tag{1.1}$$

donde a, b, s y t son funciones conocidas. La ecuación (1.1) podemos reescribirla como

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \tag{1.2}$$

donde $s(y) = \exp d(y)$ y $t(\theta) = \exp c(\theta)$. Si $a(y) = y$ se dice que la distribución es de forma canónica y algunas veces a $b(\theta)$ se le llama parámetro natural de la distribución. Si hay otros parámetros, además del parámetro de interés θ, se consideran

como parámetros de perturbación que forman parte de las funciones a, b, c, d, y se consideran conocidos.

La definición anterior puede extenderse al caso de un parámetro vectorial $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$. En tal caso, se dice que una distribución pertenece a la familia exponencial cuando su función de densidad puede expresarse de la forma

$$f(y; \boldsymbol{\theta}) = \exp[a(y)^T b(\boldsymbol{\theta}) + c(\boldsymbol{\theta}) + d(y)] \tag{1.3}$$

donde $a(y) = (a_1(y), a_2(y), ..., a_p(y))$ y $b(\boldsymbol{\theta}) = (b_1(\boldsymbol{\theta}), b_2(\boldsymbol{\theta}), ..., b_p(\boldsymbol{\theta}))$ (con a_i linealmente independientes).

1.1.1. Ejemplos de distribuciones que pertenecen a la familia exponencial

(a) La distribución exponencial. Si una variable aleatoria Y se distribuye como una exponencial, denotada por $Y \sim Exp(\lambda)$, entonces su función de densidad es

$$\begin{aligned} f(y; \lambda) &= \lambda \exp(-\lambda y) \\ &= \exp(-y\lambda + \ log\lambda) \end{aligned}$$

de ahí se tiene que $a(y) = -y$, $b(\lambda) = \lambda$, $c(\lambda) = \log \lambda$ y $d(y) = 0$, por lo tanto se concluye que la distribución exponencial pertenece a la familia exponencial.

(b) La distribución binomial. Si una variable aleatoria Y se distribuye como una binomial, denotada por $Y \sim Bin(n, \pi)$, su función de distribución de probabilidad es

$$\begin{aligned} f(y; \pi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ &= \exp\left(\log \binom{n}{y} + y \log \pi + (n-y) \log(1-\pi) \right) \\ &= \exp\left(y \log\left(\frac{\pi}{1-\pi} \right) + n \log(1-\pi) + \log \binom{n}{y} \right) \end{aligned}$$

de ahí se tiene que $a(y) = y$, $b(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, $c(\pi) = n \log(1-\pi)$ y $d(y) = \log \binom{n}{y}$, por lo tanto se concluye que la distribución binomial pertenece a la familia exponencial.

(c) La distribución Poisson. Si una variable aleatoria Y se distribuye como una Poisson, denotada por $Y \sim Pois(\lambda)$, su función de distribución de probabilidad es

$$\begin{aligned} f(y; \lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \exp(y \log \lambda - \lambda - \log y!) \end{aligned}$$

de ahí se tiene que $a(y) = y$, $b(\lambda) = \log \lambda$, $c(\lambda) = -\lambda$ y $d(y) = \log y!$, por lo tanto se concluye que la distribución poisson pertenece a la familia exponencial.

(d) La distribución Binomial Negativa. Si una variable aleatoria Y se distribuye como una Binomial Negativa, denotada por $Y \sim BinNeg(\pi, k)$, su función de distribución de probabilidad es

$$\begin{aligned} f(y;\pi,k) &= \binom{y-1}{k-1}\pi^k(1-\pi)^{y-k} \\ &= \exp\left(\log\binom{y-1}{k-1} + k\log\pi + (y-k)\log(1-\pi)\right) \\ &= \exp\left(y\log(1-\pi) + k\log\left(\frac{\pi}{1-\pi}\right) + \log\binom{y-1}{k-1}\right) \end{aligned}$$

de ahí se tiene que $a(y) = y$, $b(\pi) = \log(1-\pi)$, $c(\pi) = k\log\left(\frac{\pi}{1-\pi}\right)$ y $d(y) = \log\binom{y-1}{k-1}$, por lo tanto se concluye que la distribución binomial negativa pertenece a la familia exponencial.

(e) La distribución Normal. Si una variable aleatoria Y se distribuye como una Normal $Y \sim N(\mu, \sigma^2)$, sea $\boldsymbol{\theta} = (\mu, \sigma)$, su función de densidad es

$$\begin{aligned} f(y;\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)\right) \end{aligned}$$

de ahí se tiene que $a(y) = (a_1(y), a_2(y)) = (y^2, y)$, $b(\boldsymbol{\theta}) = (b_1(\boldsymbol{\theta}), b_2(\boldsymbol{\theta})) = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$, $c(\boldsymbol{\theta}) = -\frac{\mu^2}{2\sigma^2} + \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)$ y $d(y) = 0$, por lo tanto se concluye que la distribución normal pertenece a la familia exponencial.

(e) La distribución Gamma. Si una variable aleatoria Y se distribuye como una Gamma $Y \sim Gam(\alpha, \beta)$, sea $\boldsymbol{\theta} = (\alpha, \beta)$, su función de densidad es

$$\begin{aligned} f(y;\boldsymbol{\theta}) &= \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}\exp(-\beta y) \\ &= \exp\left(-\beta y + (\alpha-1)\log y + \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)\right) \end{aligned}$$

de ahí se tiene que $a(y) = (a_1(y), a_2(y)) = (y, \log y)$, $b(\boldsymbol{\theta}) = (b_1(\boldsymbol{\theta}), b_2(\boldsymbol{\theta})) = (-\beta, \alpha - 1)$, $c(\boldsymbol{\theta}) = \log y + \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)$ y $d(y) = 0$, por lo tanto se concluye que la distribución gamma pertenece a la familia exponencial.

1.1.2. Propiedades de la Familia Exponencial

De la definición de función de densidad de probabilidad se tiene que para todos los posibles valores de y (si la variable aleatoria Y es discreta, entonces la integral

se reemplaza por una sumatoria).

$$\int f(y;\theta)dy = 1 \tag{1.4}$$

Si derivamos en ambos lados la expresión (1.4) con respecto a θ, obtenemos

$$\frac{d}{d\theta}\int f(y;\theta)dy = \frac{d}{d\theta}1 = 0$$

podemos intercambiar el orden de integración y diferenciación, entonces se tiene

$$\int \frac{df(y;\theta)}{d\theta}dy = 0 \tag{1.5}$$

Ahora, derivamos por segunda vez (1.4) con respecto a θ e intercambiamos el orden de integración y diferenciación así obtenemos

$$\int \frac{d^2f(y;\theta)}{d\theta^2}dy = 0 \tag{1.6}$$

Aplicamos estos resultados para las distribuciones de la familia exponencial,

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

la primera derivada es

$$\frac{df(y;\theta)}{d\theta} = [a(y)b'(\theta)) + c'(\theta)]\, f(y;\theta)$$

Entonces, por (1.5) tenemos

$$\begin{aligned} 0 &= \int [a(y)b'(\theta)) + c'(\theta)]\, f(y;\theta)dy \\ &= b'(\theta)\int a(y)f(y;\theta)dy + c'(\theta)\int f(y;\theta)dy \\ &= b'(\theta)\mathbb{E}[a(y)] + c'(\theta). \end{aligned}$$

lo que implica

$$E[a(Y)] = \frac{-c'(\theta)}{b'(\theta)}. \tag{1.7}$$

Similarmente para obtener $Var[a(Y)]$ tenemos

$$\begin{aligned} \frac{d^2f(y;\theta)}{d\theta^2} &= [a(y)b''(\theta) + c''(\theta)]f(y;\theta) + [a(y)b'(\theta) + c'(\theta)]^2 f(y;\theta) \\ &= [a(y)b''(\theta) + c''(\theta)]f(y;\theta) + [b'(\theta)]^2\,[a(y) - \mathbb{E}[a(Y)]]^2\, f(y;\theta) \\ &= a(y)b''(\theta)f(y;\theta) + c''(\theta)f(y;\theta) + [b'(\theta)]^2\,[a(y) - \mathbb{E}[a(Y)]]^2\, f(y;\theta) \end{aligned}$$

por (1.5)

$$\int \frac{d^2f(y;\theta)}{d\theta^2}dy = b''(\theta)\mathbb{E}(a(Y)) + c''(\theta) + [b'(\theta)]^2 Var(a(Y)) = 0$$

lo cual implica

$$Var(a(Y)) = \frac{-b''(\theta)\mathbb{E}(a(Y)) - c''(\theta)}{[b'(\theta)]^2}$$

sustituyendo (1.6) se tiene

$$Var(a(Y)) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \tag{1.8}$$

Ahora consideremos a $\gamma = b(\theta)$ y b es una función invertible (entonces tiene una correspondencia 1-1 entre el espacio que contiene a θ y el espacio que contiene a γ), por lo que podemos volver a reescribir (1.2) como

$$f(y;\gamma) = \exp(a(y)\gamma + \kappa(\gamma) + d(y)),$$

donde $\kappa(\gamma) = c(b^{-1}(\gamma))$, por lo que se tiene el siguiente resultado.

Lema 1.1 *Supongamos que Y es una variable aleatoria cuya función de distribución pertenece a la familia exponencial canónica. Entonces, la Función Generadora de Momentos (FGM) de Y es $\mathbb{E}(exp(Yt)) = \exp(\kappa(t+\gamma) - \kappa(\gamma))$. Además, $\mathbb{E}(Y) = \kappa'(\gamma)$ y $Var(Y) = \kappa''(\gamma)$.*

Demostración: Supongamos que t es suficientemente pequeño tal que $f(y;(\theta+t))$ es una distribución. La FGM es

$$\begin{aligned} M_Y(t) = \mathbb{E}(\exp(tY)) &= \int \exp(ty)\exp(y\gamma + \kappa(\gamma) + d(y))dy \\ &= \exp(-\kappa(\gamma+t) + \kappa(\gamma)) \int \exp(y(\gamma+t) + \kappa(\gamma+t) + d(y))dy \\ &= \exp(-\kappa(\gamma+t) + \kappa(\gamma)), \end{aligned}$$

ya que $\int \exp(y(\gamma+t) + \kappa(\gamma+t) + d(y))dy = \int f(y;(\gamma+t))dy = 1$. Para obtener los momentos hay que recordar que $M_Y'(0) = \mathbb{E}(Y)$ y $Var(Y) = M_Y''(0) - (M_Y'(0))^2$. Por lo tanto

$$\begin{aligned} M_Y'(t) &= -\kappa'(\gamma+t)\exp(-\kappa(\gamma+t) + \kappa(\gamma)) \\ M_Y''(t) &= (-\kappa''(\gamma+t) + (\kappa'(\gamma+t))^2)\exp(\kappa(\gamma+t) + \kappa(\gamma)). \end{aligned}$$

De aquí se tiene que $M_Y'(0) = -\kappa'(\theta)$ y $M_Y''(0) = -\kappa''(\gamma) + \kappa'(\gamma)^2$, por lo tanto $\mathbb{E}(Y) = -\kappa'(\gamma)$ y $Var(Y) = -\kappa''(\gamma)$. ■

Ilustraremos el lema considerando una variable aleatoria $Y \sim Bin(n,\pi)$, cuya función de distribución pertenece a la familia exponencial canónica y se expresa de la siguiente forma

$$f(y;\pi) = \exp\left(y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}\right)$$

con $\gamma = b(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, $c(\pi) = n\log(1-\pi) \Longrightarrow \quad e^{\gamma} = \frac{\pi}{1-\pi}$

$$\begin{aligned} \Rightarrow \qquad & \pi = e^{\gamma}(1-\pi) \\ \Rightarrow \qquad & \pi = e^{\gamma} - e^{\gamma}\pi \\ \Rightarrow \qquad & \pi(1+e^{\gamma}) = e^{\gamma} \\ \Rightarrow \qquad & \pi = \frac{e^{\gamma}}{1+e^{\gamma}} \end{aligned}$$

por lo tanto $\kappa(\gamma) = n\log\left(1 - \frac{e^{\gamma}}{1+e^{\gamma}}\right) = n\log\left(\frac{1}{1+e^{\gamma}}\right) = n\left(\log(1) - \log(1+e^{\gamma})\right) = -n\log(1+e^{\gamma})$. Ahora encontremos el $\mathbb{E}(Y)$ y $Var(Y)$, por lo tanto

$$\begin{aligned} \mathbb{E}(Y) = -\kappa'(\gamma) &= -\left(-n\left(\frac{e^{\gamma}}{1+e^{\gamma}}\right)\right) \\ &= n\pi. \end{aligned}$$

$$\begin{aligned} Var(Y) = -\kappa''(\gamma) &= n\left(e^{\gamma}(1+e^{\gamma})^{-1} - e^{\gamma}(1+e^{\gamma})^{-2}e^{\gamma}\right) \\ &= n\left(\frac{e^{\gamma}}{1+e^{\gamma}} - \frac{e^{2\gamma}}{(1+e^{\gamma})^2}\right) \\ &= n\left(\pi - \pi^2\right) \\ &= n\pi(1-\pi). \end{aligned}$$

Una propiedad más de la familia exponencial, es que admite un estadístico suficiente para θ, para verificar este hecho recordemos los siguientes resultados:

Definición 1.1 *Un estadístico es una función real T de la muestra aleatoria $y_1, y_2, ...y_n$ de una población con función de densidad (función de probabilidad para el caso discreto) $f(Y, \theta)$. Un estadístico, T es **suficiente** para θ cuando la distribución de $(y_1, y_2, ...y_n)$ condicionada por $T = t$ no depende de θ.*

A pesar de la claridad del concepto, es muy difícil utilizar esta definición para decir si un estadístico es suficiente, por lo que tenemos una caracterización alternativa:

Teorema 1.1 ***(Criterio de factorización de Fisher-Neyman)***
*Sea $y_1, y_2, ...y_n$ una muestra aleatoria de una población con función de densidad (función de probabilidad para el caso discreto) $f(Y, \theta)$. Un estadístico T es **suficiente** para θ si y sólo si*

$$f(y_1, y_2, ...y_n; \boldsymbol{\theta}) = g(T(y_1, y_2, ...y_n), \boldsymbol{\theta})h(y_1, y_2, ...y_n). \tag{1.9}$$

donde g es una función no negativa, tanto del estadístico como del vector de parámetros, y h es una función no negativa exclusiva de los valores muestrales.

Proposición 1.1 *Sea $(y_1, y_2, ...y_n)$ una muestra aleatoria de una población Y con función de densidad (función de probabilidad para el caso discreto) perteneciente a*

la familia exponencial, entonces $f(Y;\theta)$ admite un estadístico suficiente.

Demostración: Como $f(Y;\theta)$ pertenece a la familia exponencial, entonces la función de densidad conjunta está dada por:

$$\begin{aligned} f(y_1, y_2, ...y_n;\theta) &= \prod_{i=1}^{n} f(y_i;\theta) = \prod_{i=1}^{n} \exp[a(y_i)b(\theta) + c(\theta) + d(y_i)] \\ &= \exp\left[b(\theta)\sum_{i=1}^{n} a(y_i) + nc(\theta) + \sum_{i=1}^{n} d(y_i)\right] \\ &= \exp\left[nc(\theta) + b(\theta)\sum_{i=1}^{n} a(y_i)\right] \exp\left[\sum_{i=1}^{n} d(y_i)\right] \\ &= g\left(T(y_1, y_2, ...y_n);\theta\right) h(y_1, y_2, ...y_n). \end{aligned}$$

donde $T = T(y_1, y_2, ...y_n) = \sum_{i=1}^{n} a(y_i)$, por lo tanto T es un estadístico suficiente para $f(Y;\theta)$. ■

1.2. Modelos lineales generalizados

En un modelo lineal, el valor observado de la variable dependiente Y para la observación $i-$ésima ($i = 1, 2, ..., n$), se modela como una función lineal de $(p-1)$ variables independientes $x_1, x_2, ..., x_{p-1}$ de la siguiente forma

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} + \epsilon_i \tag{1.10}$$

o de forma matricial se tiene

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{1.11}$$

donde

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

es el vector de variables dependientes, y

$$X = \begin{pmatrix} \boldsymbol{x_1^T} \\ \boldsymbol{x_2^T} \\ \vdots \\ \boldsymbol{x_n^T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \ddots & \\ 1 & x_{n1} & \cdots & x_{n(p-1)} \end{pmatrix}$$

es una matriz conocida de dimensión $n \times p$, llamada matriz diseño que contiene los valores de las variables independientes y una columna de $1's$ correspondiente a la intersección,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

que es un vector que contiene p parámetros que serán estimados y

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

es un vector de residuos. Es común suponer que los residuos en $\boldsymbol{\epsilon}$ son variables aleatorias independientes y se distribuyen normalmente, y que las varianzas son iguales **para** todos ϵ_i.

Ahora denotemos,

$$\boldsymbol{\eta} = X\boldsymbol{\beta}$$

como el predictor lineal del modelo (1.11).

Los avances en la teoría estadística y software nos permiten utilizar métodos análogos a los desarrollados para modelos lineales en las siguientes situaciones generales:

1. Las variables de respuesta tienen distribuciones distintas de la distribución normal, e incluso pueden ser categóricas en lugar de continuas.

2. La relación entre las variables de respuesta y explicativas no tiene que ser de la forma lineal simple como en (1.11).

Con todo esto, podemos decir que los **Modelos Lineales Generalizados (GLM's)** son una generalización de los modelos lineales clásicos o convencionales.

En un modelo de regresión lineal, la variable Y se distribuye normalmente, en los GLM's se permite que sigan cualquier distribución que pertenezca a la familia exponencial, además en vez de modelar $\boldsymbol{\mu} = \mathbb{E}(Y)$ directamente como una función del predictor lineal $X\boldsymbol{\beta}$ podemos modelar una función $g(\boldsymbol{\mu})$, por lo que el modelo se convierte en

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = X\boldsymbol{\beta} \tag{1.12}$$

A la función $g(\cdot)$, se le llama función liga.

1.2.1. La función liga $g(\cdot)$

La función liga $g(\cdot)$ debe cumplir dos condiciones: ser monótona y diferenciable, por lo que existe la función inversa $g(\cdot)^{-1}$.

Algunas funciones liga de uso común y sus inversas son:

- **La liga identidad:** $\boldsymbol{\eta} = \boldsymbol{\mu}$. La inversa es simplemente $\boldsymbol{\mu} = \boldsymbol{\eta}$.
- **La liga logit:** $\boldsymbol{\eta} = \log[\boldsymbol{\mu}/(1-\boldsymbol{\mu})]$. La inversa $\boldsymbol{\mu} = \frac{\exp(\boldsymbol{\eta})}{1+\exp(\boldsymbol{\eta})}$ es restringida al intervalo [0,1].
- **La liga probit:** $\boldsymbol{\eta} = \Phi^{-1}(\boldsymbol{\mu})$, donde Φ es la función de distribución Normal Estandar. La inversa $\boldsymbol{\mu} = \Phi(\boldsymbol{\eta})$ es restringida al intervalo [0,1].
- **La liga complemento log-log:** $\boldsymbol{\eta} = \log[-\log(1-\boldsymbol{\mu})]$. La inversa $\boldsymbol{\mu} = 1 - \exp(-\exp(\boldsymbol{\eta}))$ es restringida al intervalo [0,1].

Ligas canónicas

Ciertas funciones liga son comunes para ciertas distribuciones, a estas ligas se les llaman ligas canónicas. Esto significa que la liga canónica es la función $g(\cdot)$ para la cual $g(\mu) = \theta$, por ejemplo:

- **Poisson**: $\theta = \log(\mu)$ así su liga canónica es log.
- **Binomial**: $\theta = \log \frac{\pi}{1-\pi}$ así su función liga canónica es la liga logit.
- **Normal**: $\theta = \mu$ así su liga canónica es la liga identidad.

Cabe señalar que no hay garantía de que las ligas canónicas proporcionen el mejor modelo para un determinado conjunto de datos. En cualquier aplicación particular, los datos pueden mostrar un comportamiento peculiar, o puede haber justificación teórica para la elección de ligas distintas de las ligas canónicas [4].

1.2.2. La devianza

Sea Y_i la variable de respuesta asociada al vector con $p-1$ variables explicativas $\boldsymbol{X}_i = (x_{i1}, x_{i2}, ..., x_{i(p-1)})$. El modelo $g(\mathbb{E}(Y_i)) = \boldsymbol{x_i^T \beta}$ que considera a todas las variables explicativas se conoce como modelo saturado, cuando consideramos solo un subconjunto de la variables explicativas (subconjunto de las coordenadas de $\boldsymbol{X}_i$), $\boldsymbol{X}_i^* = (x_{i1}^*, x_{i2}^*, ..., x_{im}^*)$ con $0 < m < p-1$ el modelo que considera solo a este conjunto de variables explicativas $g(\mathbb{E}(Y_i)) = \boldsymbol{x_i^{*T} \beta^*}$ se conoce como modelo de interés.

Sea $\widehat{\boldsymbol{\beta}}$ el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ en el modelo saturado y $\widehat{\boldsymbol{\beta}}^*$ el estimador de máxima verosimilitud de $\boldsymbol{\beta^*}$ en el modelo de interés, entonces, la función de verosimilitud $L(\widehat{\boldsymbol{\beta}}; y)$, es mayor o igual a $L(\widehat{\boldsymbol{\beta}}^*; y)$, con la misma distribución y la misma función liga.

Para determinar que variables explicativas influyen en el resultado de la variable Y_i, se deben comparar el modelo saturado con el modelo de interés mediante un término conocido como devianza.

Definición 1.2 *Sea $l(\widehat{\boldsymbol{\beta}}^*; y)$ la función log-verosimilitud para el modelo de interés y $l(\widehat{\boldsymbol{\beta}}; y)$ es la función log-verosimilitud para el modelo saturado. La devianza D está dada por:*

$$D = 2[l(\widehat{\boldsymbol{\beta}}; y) - l(\widehat{\boldsymbol{\beta}}^*; y)]. \tag{1.13}$$

Si el modelo de interes es adecuado, la devianza tendrá una distribución asintótica χ^2 con $p-m$ grados de libertad a medida que aumenta n. Esto puede ser utilizado como una prueba de la adecuación del modelo. El grado de aproximación de la devianza a la distribución χ^2 depende del tipo de distribución que se trate.

Un segundo, y quizás más importante uso de la devianza es en la comparación de modelos de competencia. Supongamos que un determinado modelo tiene una devianza D_1 con gl_1 grados de libertad, y que un modelo más simple produce una

devianza D_2 con gl_2 grados de libertad. El modelo más simple tendrá una devianza mayor y más grados de libertad (gl), entonces, para comparar los dos modelos se puede calcular la diferencia de las devianzas ($D_2 - D_1$), y se relaciona esto con una distribución χ^2 con ($gl_2 - gl_1$) grados de libertad. Esto daría una prueba de la importancia de los parámetros que se incluyen en el modelo 1 pero no en el modelo 2. Esto, por supuesto, requiere que los parámetros incluidos en el modelo 2 sea un subconjunto de la parámetros del modelo 1, es decir, que se anidan los modelos.

1.2.3. Residuales

El residual es la diferencia entre el valor observado y y el valor estimado $\widehat{y}$, esto es $\widehat{e} = y - \widehat{y}$. Si al ajustar el modelo, los residuos resultantes son pequeños, se tendría evidencia que el modelo es adecuado. Por esta razón se utiliza el análisis de residuales como criterio para determinar si un conjunto de datos se ajusta al modelo propuesto.

Residual de un modelo lineal

El estimador del valor esperado del vector y, o valor ajustado y es una función lineal de los valores observados.

$$\widehat{E(y)} = \widehat{y} = \mathbf{Hy}$$

donde **H** es conocida como la matriz sombrero y se tiene que ésta matriz es idempotente y simétrica.

Ejemplo: En regresión lineal tenemos $\widehat{y} = X\widehat{\beta} = X(X^TX)^{-1}X^Ty$. En este caso, la matriz sombrero es $H = X(X^TX)^{-1}X^T$.

La matriz sombrero puede tener una forma más compleja en los modelos lineales generalizados, pero se tiene que la mayoría del software tiene opciones para imprimir la matriz sombrero.

Residuos de Pearson

El residual para una observación y_i se puede definir como $\widehat{e}_i = y_i - \widehat{y}_i$. El residual de Pearson es el residual estandarizado con la desviación estándar del valor ajustado:

$$e_{i,Pearson} = \frac{y_i - \widehat{y}_i}{\sqrt{\widehat{Var}(\widehat{y}_i)}}$$

Los residuos de Pearson tienen la característica de que tienden asintóticamente a una distribución normal con varianza constante. Sin embargo, la varianza de los residuos de Pearson no necesariamente es 1. El error estándar de los residuos de Pearson es:

$$e_{i,adj,P} = \frac{e_{i,Pearson}}{\sqrt{1 - h_{ii}}}$$

donde h_{ii} son los elementos de la diagonal de la matriz sombrero.

Residuos de la Devianza

La devianza D puede escribirse como una suma $D = \sum_i d_i$ donde las d_i son los que se conocen como componentes de la devianza que vienen de la ecuación de verosimilitud y de su derivada, cada observación i contribuye una cantidad d_i, y se definen los residuos de la devianza como

$$e_{i,Devianza} = sign(y_i - \widehat{y}_i)\sqrt{d_i}$$

Los residuos de la devianza también se pueden escribir en forma estandarizada, es decir, tal que su varianza sea cercana a la unidad. Esto se obtiene como

$$e_{i,adj,D} = \frac{e_{i,Devianza}}{\sqrt{1 - h_{ii}}}$$

Residuos Score

El estimador de máxima verosimilitud del parámetro θ, es la solución de la ecuación de puntuación

$$U = \frac{\partial l}{\partial \theta} = 0$$

Esta ecuación, para la familia exponencial corresponde a una suma de términos U_i para cada observación, donde las U_i son las derivadas parciales de la logverosimilitud del modelo con respecto a cada parámetro θ_i. Estos términos pueden ser debidamente estandarizado e interpretados como residuos, es decir, como la contribución de cada observación al punto. Esta estandarización de residuos puntuales son obtenidos de

$$e_{i.adj,S} = \frac{U_i}{\sqrt{(1 - h_{ii})v_i}}$$

donde h_{ii} son elementos diagonales de la matriz estimada y v_i son elementos de una cierta matriz de pesos.

Residuos de Verosimilitud

Teóricamente sería posible comparar la devianza de un modelo que comprende todos los datos con la devianza de un modelo con la observación i excluida. Una aproximación a los residuos que serían obtenidos usando este procedimiento es

$$e_{i,Probabilidad} = sign(y_i - \widehat{y}_i)\sqrt{h_{ii}(e_{i,Score})^2 + (1 - h_{ii})(e_{i,Devianza})^2}$$

que es un promedio ponderado de los residuos de la devianza y de los residuos score.

1.2.4. Modelo Poisson

Sean las variables de respuesta $Y_1, ..., Y_N$ independientes y $Y_i \sim Poisson(\lambda_i)$ esto es,

$$f(y_i, \lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

y sea $\boldsymbol{X}_i = (x_{i1}, x_{i2}, ..., x_{i(p-1)})$ el conjunto de variables explicativas, entonces se tiene que $\mathbb{E}(Y_i|\boldsymbol{x}_i) = \lambda_i = Var(Y_i|\boldsymbol{x}_i)$ y

$$g(\lambda_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i(p-1)}$$

utilizando la liga log tenemos que $g(\lambda_i) = \log(\lambda_i)$

$$\Rightarrow \qquad \lambda_i = e^{\boldsymbol{x_i^T \beta}} \tag{1.14}$$

la función de verosimilitud está dada por:

$$L(\boldsymbol{\lambda}, y) = L(\lambda_1, ..., \lambda_N; y_1, ..., y_N) = \prod_{i=1}^{N} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_1!} \tag{1.15}$$

por lo tanto, la función log-verosimilitud está dada por:

$$\begin{aligned} l(\boldsymbol{\lambda}; y) &= \sum_{i=1}^{N} y_i \log \lambda_i - \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \log y_i! \\ &= \sum_{i=1}^{N} y_i \log(e^{\boldsymbol{x_i^T \beta}}) - \sum_{i=1}^{N} e^{\boldsymbol{x_i^T \beta}} - \sum_{i=1}^{N} \log y_i! \\ &= \sum_{i=1}^{N} y_i \boldsymbol{x_i^T \beta} - \sum_{i=1}^{N} e^{\boldsymbol{x_i^T \beta}} - \sum_{i=1}^{N} \log y_i! \\ &= \sum_{i=1}^{N} y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i(p-1)}) \\ &\quad - \sum_{i=1}^{N} e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i(p-1)}} - \sum_{i=1}^{N} \log y_i! \end{aligned} \tag{1.16}$$

ahora, derivando e igualando a cero se tiene

$$\frac{\partial l(\boldsymbol{\lambda}; y)}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} e^{\boldsymbol{x_i^T \beta}} x_{ij} = 0$$

Lo cual implica

$$\sum_{i=1}^{N} y_i x_{ij} = \sum_{i=1}^{N} e^{\boldsymbol{x_i^T \beta}} x_{ij}. \tag{1.17}$$

Para obtener el vector máximo verosímil $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_{p-1})$ tenemos que resolver la ecuación (1.17) mediante métodos numéricos de aproximación.

Sea $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})$ el vector de parámetros del modelo saturado y sea $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_{p-1})$ el estimador máximo verosímil del vector de parámetros, entonces sustituyendo en (1.17) se tiene que:

$$l(\widehat{\boldsymbol{\beta}}; y) = \sum_{i=1}^{N} y_i \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} - \sum_{i=1}^{N} e^{\boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}} - \sum_{i=1}^{N} \log y_i! \tag{1.18}$$

donde $\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} = \hat{\beta}_0 + \widehat{\beta}_1 x_{i1} + ... + \widehat{\beta}_{p-1} x_{i(p-1)}$.

Ahora sea $\boldsymbol{X}_i^* = (x_{i1}^*, x_{i2}^*, ..., x_{im}^*)$ un subconjunto de variables explicativas de $\boldsymbol{X}_i$ con $0 < m < p-1$, entonces sea $\boldsymbol{\beta^*} = (\beta_0^*, \beta_1^*, ..., \beta_m^*)$ el vector de parámetros del modelo de interés y sea $\widehat{\boldsymbol{\beta^*}} = (\widehat{\beta}_0^*, \widehat{\beta}_1^*, ..., \widehat{\beta}_m^*)$ el estimador máximo verosímil del vector de parámetros, entonces sustituyendo en (1.14) se tiene que:

$$l(\widehat{\boldsymbol{\beta^*}}; y) = \sum_{i=1}^{N} y_i \boldsymbol{x}_i^{*T} \widehat{\boldsymbol{\beta^*}} - \sum_{i=1}^{N} e^{\boldsymbol{x}_i^{*T}\widehat{\boldsymbol{\beta}}^*} - \sum_{i=1}^{N} \log y_i! \tag{1.19}$$

donde $\boldsymbol{x}_i^{*T} \widehat{\boldsymbol{\beta^*}} = \widehat{\beta}_0^* + \widehat{\beta}_1^* x_{i1}^* + ... + \widehat{\beta}_m^* x_{im}^*$.

Así, sustituyendo (1.18) y (1.19) en (1.13) tenemos que la devianza para un modelo Poisson está dada por:

$$\begin{aligned} D &= \left(\sum_{i=1}^{N} y_i \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} - \sum_{i=1}^{N} e^{\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}} - \sum_{i=1}^{N} \log y_i! \right) \\ &\quad - \left(\sum_{i=1}^{N} y_i \boldsymbol{x}_i^{*T} \widehat{\boldsymbol{\beta^*}} - \sum_{i=1}^{N} e^{\boldsymbol{x}_i^{*T}\widehat{\boldsymbol{\beta}}^*} - \sum_{i=1}^{N} \log y_i! \right) \\ &= \sum_{i=1}^{N} y_i \left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} - \boldsymbol{x}_i^{*T} \widehat{\boldsymbol{\beta^*}} \right) - \sum_{i=1}^{N} \left(e^{\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}} - e^{\boldsymbol{x}_i^{*T}\widehat{\boldsymbol{\beta}}^*} \right). \end{aligned}$$

1.2.5. Modelo Binomial Negativa

Si las variables de respuesta $Y_1, ..., Y_N$ son independientes y $Y_i \sim BinNeg(\pi_i, k)$ y sea $\boldsymbol{X}_i = (x_{i1}, x_{i2}, ..., x_{i(p-1)})$ el conjunto de variables explicativas, entonces se tiene que $\mathbb{E}(Y_i|\boldsymbol{x}_i) = \mu_i = \frac{k}{\pi_i} \quad \Rightarrow \quad \pi_i = \frac{k}{\mu_i}$ y su función de probabilidad está dada por:

$$\begin{aligned} f(y_i; \pi_i, k) &= \binom{y_i - 1}{k-1} \pi_i^k (1-\pi_i)^{y_i - k} \\ &= \binom{y_i - 1}{k-1} \left(\frac{k}{\mu_i} \right)^k \left(1 - \frac{k}{\mu_i} \right)^{y_i - k} \\ &= \binom{y_i - 1}{k-1} \left(\frac{k}{\mu_i} \right)^k \left(\frac{\mu_i - k}{\mu_i} \right)^{y_i - k} \end{aligned}$$

y tenemos

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{p-1} x_{i(p-1)}$$

utilizando la liga log tenemos que $g(\mu_i) = \log(\mu_i)$

$$\Rightarrow \qquad \mu_i = e^{\boldsymbol{x_i^T \beta}} \tag{1.20}$$

así la función de verosimilitud está dada por:

$$L(\boldsymbol{\pi}, y) = \prod_{i=1}^{N} \left[\binom{y_i - 1}{k-1} \left(\frac{k}{\mu_i} \right)^k \left(\frac{\mu_i - k}{\mu_i} \right)^{y_i - k} \right] \tag{1.21}$$

por lo tanto, la función log-verosimilitud está dada por:

$$\begin{aligned}
l(\boldsymbol{\pi}; y) &= \sum_{i=1}^{N} \left[\log \binom{y_i - 1}{k-1} + k \log \left(\frac{k}{\mu_i} \right) + (y_i - k) \log \left(\frac{\mu_i - k}{\mu_i} \right) \right] \\
&= \sum_{i=1}^{N} \left[\log \binom{y_i - 1}{k-1} + k \left(\log k - \log \mu_i \right) + (y_i - k) \left[\log(\mu_i - k) - \log \mu_i \right] \right] \\
&= \sum_{i=1}^{N} \left[C_i - k \log \mu_i + (y_i - k) \log(\mu_i - k) - (y_i - k) \log \mu_i \right] \\
&= \sum_{i=1}^{N} \left[C_i - y_i \log \mu_i + (y_i - k) \log(\mu_i - k) \right] \\
&= \sum_{i=1}^{N} \left[C_i - y_i \boldsymbol{x_i^T \beta} + (y_i - k) \log(e^{\boldsymbol{x_i^T \beta}} - k) \right]
\end{aligned}$$

donde $C_i = \left[\log \binom{y_i - 1}{k-1} + k \log k \right]$ es una constante que no depende de μ_i, ahora, derivando e igualando a cero se tiene

$$\frac{\partial l(\boldsymbol{\pi}; y)}{\partial \beta_j} = \sum_{i=1}^{N} \left[-y_i x_{ij} + \frac{(y_i - k) e^{\boldsymbol{x_i^T \beta}} x_{ij}}{(e^{\boldsymbol{x_i^T \beta}} - k)} \right] = 0$$

Lo cual implica

$$\sum_{i=1}^{N} \left[\frac{(y_i - k) e^{\boldsymbol{x_i^T \beta}} x_{ij}}{(e^{\boldsymbol{x_i^T \beta}} - k)} \right] = \sum_{i=1}^{N} y_i x_{ij} \tag{1.22}$$

Para obtener el vector máximo verosímil $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_{p-1})$ tenemos que resolver la ecuación (1.22) mediante métodos numéricos de aproximación.

1.3. Análisis de conglomerados (Clústers)

El objetivo de este análisis es formar grupos de observaciones, de manera que todas las unidades en un grupo sean lo más homogéneas posibles entre ellas pero que sean heterogéneas a aquellas de otros grupos [11]. En el análisis de conglomerados existen métodos jerárquicos y no- jerárquicos y se estudian tres tipos de problemas:

1. **Partición de los datos**. Disponemos de datos que sospechamos son heterogéneos y se desea separarlos en un número de grupos prefijado, de manera que:

   a) cada elemento pertenezca a uno y solo uno de los grupos;

   b) todo elemento quede clasificado;

   c) cada grupo sea internamente homogéneo.

2. **Construcción de jerarquías.** Deseamos estructurar elementos de un conjunto de forma jerárquica por su similitud.

3. **Clasificación de variables.** En problemas con muchas variables es necesario hacer un estudio exploratorio inicial para dividir las variables en grupos.

1.3.1. Distancias y Disimilitudes

Para entrar al tema de análisis de conglomerados es muy importante introducir los conceptos de distancia y disimilitud.

Para un conjunto de elementos de $\mathcal{A}$ se define **distancia** entre dos elementos a y b del conjunto como una función matemática $d(a, b)$ que cumple las siguientes propiedades:

1. $d(x, y) \geq 0 \quad \forall x, y \in \mathcal{A}$
2. $d(x, y) = d(y, x) \quad \forall x, y \in \mathcal{A}$
3. $d(x, y) = 0 \quad ssi \quad x = y \quad \forall x, y \in \mathcal{A}$
4. $d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z \in \mathcal{A}$

Una **disimilitud** d^* es parecida a una distancia solo que no cumplen todas las propiedades;

1. $d^*(x, y) \geq 0 \quad \forall x, y \in \mathcal{A}$ (acotada por arriba)
2. $d^*(x, y) = d^*(y, x) \quad \forall x, y \in \mathcal{A}$
3. $d^*(x, x) =$ínfimo (acotada por abajo)

En el análisis de conglomerados podemos tener observaciones con diferentes tipos de datos, es por eso que se tienen diferentes tipos de distancias y disimilitudes.

Distancias entre individuos con datos numéricos

En estos tipos de datos la distancia euclidiana puede generalizarse introduciendo una matriz A definida positiva de dimensiones $p \times p$ como sigue:

$$d(x, y)^2 = ((x - y)^T A(x - y)$$

cuando $A = S^{-1}$ es el inverso de la matriz de varianzas y covarianzas se le conoce como distancia de *Mahalanobis*.
Otra distancia es la *city block*, corresponde a:

$$d(x, y) = |(x_1 - y_1)| + ... + |(x_p - y_p)|$$

Esta distancia también puede generalizarse como:

$$d(x, y) = [(x_1 - y_1)^\alpha + ... + (x_p - y_p)^\alpha]^{1/\alpha}$$

con α un entero, y es conocida como distancia de *Minkowski*.

Distancias entre individuos con datos de distintas escalas

Cuando se consideran datos con varios tipos de escalas de medición, se utiliza el **Índice de Gower** para crear disimilitudes. Primero se crea una variable δ que permita decir si son comparables o no los sujetos en esa variable:

$$\delta_{ijk} = \begin{cases} 1 & \text{si} \quad \text{Se puede comparar a } i \text{ con } j \text{ en la variable } x_k \\ 0 & \text{si} \quad \text{No se puede comparar a } i \text{ con } j \text{ en la variable } x_k \end{cases}$$

Después para cada variable x_k se mide la similitud entre el individuo i y el j, la similitud entre i y j está dada por:

$$c_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

y la disimilitud se define como $d_{ij} = 1 - c_{ij}$. Si x_k es cuantitativa

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{rango(x_k)}$$

Si x_k es cualitativa

$$s_{ijk} = \begin{cases} 1 & \text{si} \quad i \text{ concuerda con } j \text{ en la variable } x_k \\ 0 & \text{si} \quad i \text{ no concuerda con } j \text{ en la variable } x_k \end{cases}$$

Donde s_{ijk} es el coeficiente de similaridad según la variable k entre dos elementos muestrales $(i.j)$ y se define como una función no negativa y simétrica:

(a) $s_{ijk} = 1$

(b) $0 \leq s_{ijk} \leq 1$

(c) $s_{ijk} = s_{jik}$

1.3.2. Métodos Jerárquicos

Los métodos jerárquicos tienen por objetivo agrupar conglomerados para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso se minimice alguna distancia o bien se maximice alguna medida de similitud, estos métodos se subdividen en aglomerativos y divisivos.

1. **Aglomerativos.** Parten de los elementos individuales y los van agregando en grupos, estos métodos son iterativos y en cada paso debe recalcularse la matriz de distancias.

2. **Divisivos.** Parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales.

Esta manera de formar nuevos grupos tiene la particularidad de que si en un determinado nivel se agrupan dos clusters, éstos quedan ya jerárquicamente agrupados para el resto de los niveles.

Métodos aglomerativos

En estos métodos los algoritmos que se utilizan tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos. Su estructura es:

1. Comenzar con tantas clases como elementos n. Las distancias entre clases son las distancias entre elementos originales.

2. Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.

3. Recalcular la matriz de distancias y sustituir los dos elementos utilizados en (2) para definir la clase por un nuevo elemento que represente la clase construida.

4. Volver a (2) y repetir (2) y (3) hasta que tengamos todos los elementos agrupados en una clase única.

El punto principal de este tipo de métodos es calcular las distancias entre grupos. Supongamos que tenemos un grupo C_i con n_i elementos, y un grupo C_j con n_j elementos para determinar la distancia entre ambos grupos tomaremos en cuenta los siguientes métodos:

a) *Liga sencilla o del vecino más cercano:* En este método se considera que la distancia o similitud entre dos conglomerados está dada, respectivamente, por la mínima distancia (o máxima similitud) entre sus componentes, así la distancia y similitud entre ellos será:

$$d(C_i, C_j) = \min_{x_i \in C_l, x_m \in C_j} d(x_l, x_m) \quad l = 1, ..., n_i \ ; \ m = 1, ..., n_j$$

$$s(C_i, C_j) = \max_{x_i \in C_l, x_m \in C_j} s(x_l, x_m) \quad l = 1, ..., n_i \ ; \ m = 1, ..., n_j$$

Como este criterio sólo depende del orden de las distancias (o similitudes) será invariante ante transformaciones monótonas: obtendremos la misma jerarquía aunque las distancias sean numéricamente distintas.

b) *Liga completa o del vecino más lejano:* En este método, se considera que la distancia o similitud entre dos conglomerados hay que medirla atendiendo a sus elementos más dispares, es decir, la distancia o similitud entre conglomerados viene dada, respectivamente, por la máxima distancia (o mínima similitud) entre sus componentes, así la distancia y similitud entre ellos será:

$$d(C_i, C_j) = \max_{x_i \in C_l, x_m \in C_j} d(x_l, x_m) \quad l = 1, ..., n_i \ ; \ m = 1, ..., n_j$$

$$s(C_i, C_j) = \min_{x_i \in C_l, x_m \in C_j} s(x_l, x_m) \quad l = 1, ..., n_i \ ; \ m = 1, ..., n_j$$

Este criterio será también invariante ante transformaciones monótonas de las distancias al depender, como el anterior, del orden de las distancias (o similitudes).

c) *Liga promedio:* En este método se considera que la distancia, o similitud, entre dos conglomerados, viene definida por el promedio ponderado de las distancias, o similitudes, de los componentes de un conglomerado respecto a los del otro, así su distancia y similitud entre ellos sera:

$$d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{l=1}^{n_i} \sum_{m=1}^{n_j} d(x_l, x_m)$$

$$s(C_i, C_j) = \frac{1}{n_i n_j} \sum_{l=1}^{n_i} \sum_{m=1}^{n_j} s(x_l, x_m)$$

Como se ponderan los valores de las distancias, este criterio no es invariante ante transformaciones monótonas de las distancias.

d) *Método Ward:* Este método es un procedimiento en el cual, en cada etapa, se unen los dos conglomerados para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias de cada conglomerado, de cada individuo al centroide del conglomerado, es decir:

$$SCDG = \sum_{k=1}^{n} \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{jk})^2 \tag{1.23}$$

donde x_{ijk} es el valor de la variable j en el elemento i del grupo k y $\bar{x}_{jk}$ la media de esta variable en el grupo. Este método comienza suponiendo que cada dato forma un grupo, es decir que tenemos n grupos por tanto $SCDG = 0$, después se unen los elementos que produzcan el incremento mínimo de $SCDG$, es decir, tomar los más próximos con la distancia euclidiana. En la siguiente etapa tenemos $n - 1$ grupos, $n - 2$ con un solo elemento y uno con dos elementos. De nuevo se unen dos grupos tales que $SCDG$ tenga un incremento mínimo, para obtener $n - 2$ grupos y así sucesivamente hasta obtener un único grupo.

Métodos Divisivos

Como se había descrito anteriormente, los métodos divisivos, constituyen el proceso inverso a los aglomerativos. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez menores.

Una cuestión importante que puede surgir en su desarrollo es el hecho de cuando un conglomerado determinado debe dejar de dividirse para proceder con la división de otro conglomerado distinto. Dicho procedimiento comienza con la eliminación del grupo principal de aquel individuo cuya distancia sea mayor, o cuya similaridad sea menor, al conglomerado formado por los restantes individuos. Así se tiene un conglomerado unitario y otro formado por los restantes individuos. A continuación se añadirá al conglomerado unitario aquel elemento cuya distancia (similaridad) total al resto de los elementos que componen su actual conglomerado menos la distancia (similaridad) al conglomerado anteriormente formado sea máxima (mínima). Cuando esta diferencia sea negativa dicho elemento no se añade y se repite el proceso sobre los dos subgrupos.

Dendograma

Un dendograma es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de conglomerados. Los elementos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud entre ellos, su construcción es de la siguiente manera:

(i) En la parte inferior del gráfico se disponen los n elementos iniciales.

(ii) Las uniones entre elementos se representan por tres líneas rectas. Dos dirigidas a los elementos que se unen y que perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel en que se unen.

(iii) El proceso se repite hasta que todos los elementos están conectados por líneas rectas.

Si cortamos el dendrograma a un nivel de distancia dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman. El dendrograma es útil cuando los puntos tienen claramente una estructura jerárquica, pero puede ser engañoso cuando se aplica ciegamente, ya que dos puntos pueden parecer próximos cuando no lo están, y pueden aparecer alejados cuando están próximos.

Algunos ejemplos

Supongamos que tenemos 7 elementos A,B,C,D,E,F,G cuya matriz de distancias esta dada por:

	A	B	C	D	E	F	G
A	0						
B	2.15	0					
C	0.7	1.53	0				
D	1.07	1.14	0.43	0			
E	0.85	1.38	0.21	0.29	0		
F	1.16	1.01	0.55	0.22	0.41	0	
G	1.56	2.38	1.86	2.04	2.02	2.05	0

El primer paso es tomar los dos elementos más próximos es decir, el primer grupo los formarán (C,E) cuya distancia entre ellos es de 0.21, después hay que recalcular la matriz de distancias, en este caso consideraremos **la liga sencilla**, es decir, tomaremos la mínima distancia entre (C, E) con el resto de los elementos y así la nueva matriz de distancias queda de la siguiente forma:

	A	B	(C,E)	D	F	G
A	0					
B	2.15	0				
(C,E)	0.7	1.38	0			
D	1.07	1.14	0.29	0		
F	1.16	1.01	0.41	0.22	0	
G	1.56	2.38	1.86	2.04	2.05	0

De nuevo se toman los elementos más próximos, es decir (D,F) formarán el nuevo grupo ya que la distancia entre ellos es 0.22, así recalculando la matriz de distancias se tiene:

	A	B	(C,E)	(D,F)	G
A	0				
B	2.15	0			
(C,E)	0.7	1.38	0		
(D,F)	1.07	1.01	0.29	0	
G	1.56	2.38	1.86	2.04	0

De nuevo se toman los elementos más próximos, es decir ((D,F),(C,E)) formarán el nuevo grupo, recalculando la matriz de distancias se tiene:

	A	B	((C,E),(D,F))	G
A	0			
B	2.15	0		
((C,E),(D,F))	0.7	1.01	0	
G	1.56	2.38	1.86	0

De nuevo se toman los elementos más próximos, es decir (A,((D,F),(C,E))) formarán el nuevo grupo, recalculando la matriz de distancias se tiene:

	(A,((C,E),(D,F)))	B	G
(A,((C,E),(D,F)))	0		
B	1.01	0	
G	1.56	2.38	0

De nuevo se toman los elementos más próximos, es decir (B,(A,((D,F),(C,E))))formarán el nuevo grupo, recalculando la matriz de distancias se tiene:

	(B,(A,((D,F),(C,E))))	G
(B,(A,((D,F),(C,E))))	0	
G	1.56	0

Una vez ya formados los grupos los podemos visualizar mediante el dendograma, que queda de la siguiente forma:

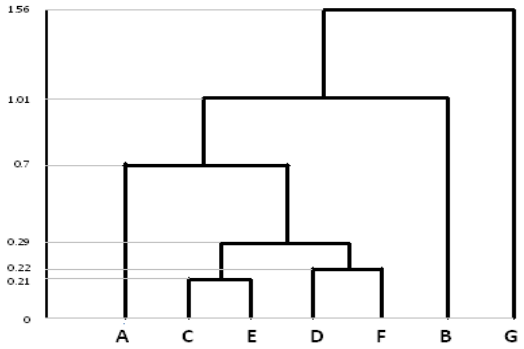
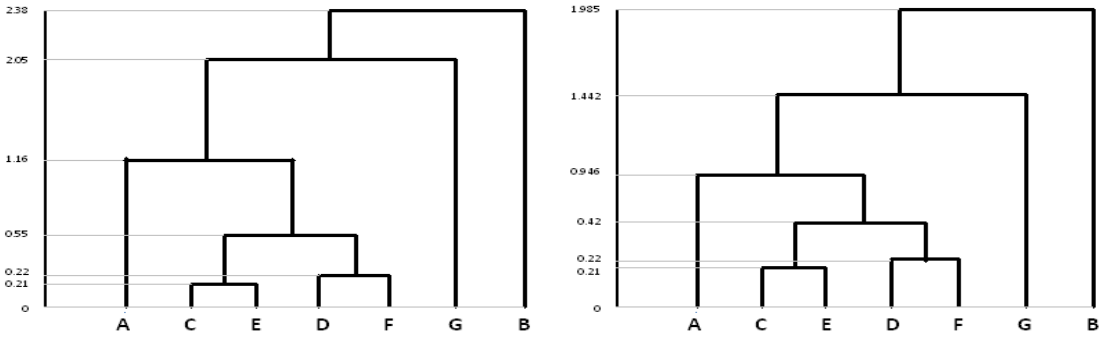


Figura 1.1: Dendograma liga sencilla

De manera análoga recalculando la matriz de distancias utilizando las ligas completas y promedio se tiene que los elementos quedan agrupados de la siguiente manera:



a) Liga completa b) Liga promedio

Dendogramas

Método Ward: En este caso consideremos 5 elementos A,B,C,D,F, con dos componentes x_1, x_2:

	x_1	x_2
A	10	5
B	20	20
C	30	10
D	30	15
E	5	10

Empezando con 5 grupos, tenemos que encontrar las posibles formas de tener 4 grupos y calcular cuál de ellos produce un incremento mínimo a (1.22) y lo resumimos en la siguiente tabla:

Grupos	Centroides	$SCDG_k$	$SCDG$
(A,B),C,D,E	$\bar{x}_{AB}$ =(15,12.5)	(A,B)=162.5 C=D=E=0	162.5
(A,C),B,D,E	$\bar{x}_{AC}$ =(20,7.5)	(A,C)=212.5 B=D=E=0	212.5
(A,D),B,C,E	$\bar{x}_{AD} = (20, 10)$	(A,D)=250 B=D=E=0	250
(A,E),B,C,D	$\bar{x}_{AE}$ =(7.5,7.5)	(A,E)=25 B=C=D=0	25
(B.C),A,D,E	$\bar{x}_{BC} = (25, 15)$	(B,C)=100 A=D=E=0	100
(B.D),A,C,E	$\bar{x}_{BD} = (25, 17{,}5)$	(B,D)=62.5 A=C=E=0	62.5
(B,E),A,C,D	$\bar{x}_{BE} = (12{,}5, 15)$	(B,E)=162.5 A=C=D=0	162.5
(C,D),A,B,E	$\bar{x}_{CD}$ =(30,12.5)	(C,D)=12.5 A=B=E=0	12.5
(C,E),A,B,D	$\bar{x}_{CE}$ =(17.5,10)	(C,E)=312.5 A=B=D=0	312.5
(D,E),A,B,C	$\bar{x}_{DE}$ =(17.5,12.5)	(D,E)=325 A=B=C=0	325

Como se puede observar, el grupo que al unirse incrementa en mínimo (1.22) son (C,D), ahora encontremos las posibilidades para tener 3 grupos:

Grupos	Centroides	$SCDG_k$	$SCDG$
(A,(C,D)),B,E	$\bar{x}_{ACD}$ =(23.33,10)	(A,(C,D))=316.66 B=E=0	316.66
(B,(C,D)),A,E	$\bar{x}_{BCD}$ =(26.66,15)	(B,(C,D))=116.66 A=E=0	116.66
((C,D),E),A,B	$\bar{x}_{CDE}$ =(21.66,11.66)	((C,D),E)=433.33 A=B=0	433.33
(A,B),(C,D),E	$\bar{x}_{AB}$ =(15,12.5) $\bar{x}_{CD}$ =(30,12.5)	(A,B)=162.5 (C,D)=12.5 E=0	175
(A,E),(C,D),B	$\bar{x}_{AE}$ =(7.5,7.5) $\bar{x}_{CD}$ =(30,12.5)	(A,E)=25 (C,D)=12.5 B=0	37.5
(B,E),(C,D),A	$\bar{x}_{BE}$ =(12.5,15) $\bar{x}_{CD}$ =(30,12.5)	(B,E)=162.5 (C,D)=12.5 A=0	175

Así tenemos que se forma un nuevo grupo con (A,E), ahora encontremos las posibles maneras de formar 2 grupos:

Grupos	Centroides	$SCDG_k$	$SCDG$
((C,D),(A,E)),B	$\bar{x}_{CDAE}$ =(18.75,10)	((C,D),(A,E))=568.75 B=0	568.75
(C,D),((A,E),B)	$\bar{x}_{CD}$ =(30,12.5) $\bar{x}_{AEB}$ =(11.66,11.66)	(C,D)=12.5 ((A,E),B)=233.33	245.8
((C,D),B),(A,E)	$\bar{x}_{CDB}$ =(26,66,15) $\bar{x}_{AE}$ =(7.5,7.5)	((C,D),B)=116.66 (A,E)=25	141.66

De aquí se observa que el elemento B forma un nuevo grupo con (C,D) y así tenemos el ultimo grupo dado por:

Grupos	Centroides	$SCDG_k$	$SCDG$
((A,E),((C,D),B))	$\bar{x}_{AECDB}$ =(19,12)	((C,D),((A,E),B)) =650	650

y si el dendograma queda de la siguiente forma:

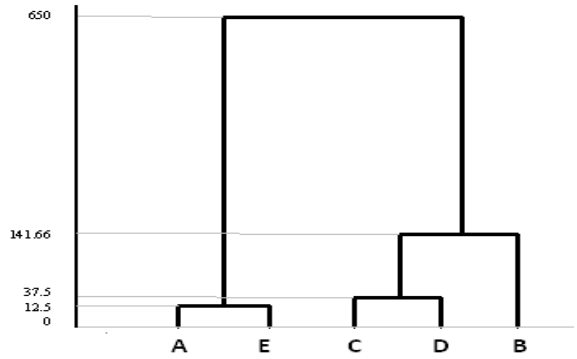


Figura 1.2: Dendograma método Ward

1.3.3. Métodos no Jerárquicos

Estos métodos que están diseñados para clasificar individuos en una clasificación de K conglomerados, donde K se determina como una parte del proceso. La idea central de la mayoría de estos procedimientos es elegir alguna partición inicial de individuos y después intercambiar los miembros de estos conglomerados para obtener una partición mejor.

Los diversos algoritmos existentes se diferencian sobre todo en lo que se entiende por una partición mejor y en los métodos que deben usarse para conseguir mejoras. Tales algoritmos empiezan con un punto inicial y generan una secuencia de movimientos de un punto a otro hasta que se encuentra un óptimo local de la función objetivo. Los métodos estudiados ahora comienzan con una partición inicial de los individuos en grupos o bien con un conjunto de puntos iniciales sobre los cuales pueden formarse los conglomerados.

Método de K-medias

Supongamos una muestra de n elementos con p variables. El objetivo es dividir esta muestra en un número de grupos prefijado, K, este algoritmo requiere de cuatro etapas:

1. Seleccionar K puntos como centros de los grupos iniciales, esto puede hacerse:

   a) Asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos así formados.

   b) Tomando como centros los K puntos más alejados entre sí.

   c) Construyendo los grupos con información a priori, o bien seleccionando los centros a priori.

2. Calcular las distancias euclidianas de cada elemento al centro de los K grupos, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.

3. Definir un criterio de optimización y comprobar si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio.

4. Si no es posible mejorar el criterio de optimización, terminar el proceso.

El criterio de homogeneidad que se utiliza en el algoritmo de k-medias es la suma de cuadrados dentro de los grupos ($SCDG$)para todas las variables, que es equivalente a la suma ponderada de las varianzas de las variables en los grupos:

$$SCDG = \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_{jk})^2$$

donde x_{ijk} es el valor de la variable j en el elemento i del grupo k y $\bar{x}_{jk}$ la media de esta variable en el grupo El criterio de optimización se escribe como:

$$\text{mín}\, SCDG = \text{mín} \sum_{k=1}^{K} \sum_{j=1}^{p} n_k v_{jk}^2 \tag{1.24}$$

donde n_k es el número de elementos del grupo k y v_{jk}^2 es la varianza de la variable j en dicho grupo. La varianza de cada variable en cada grupo es claramente una medida de heterogeneidad del grupo y al minimizar las varianzas en los grupos obtendremos grupos más homogéneos.

Un posible criterio alternativo de homogeneidad sería minimizar las distancias al cuadrado entre los centros de los grupos y los puntos que pertenecen a ese grupo. Si medimos las distancias con la norma euclidiana, este criterio se escribe como:

$$\text{mín} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^T (x_{ik} - \bar{x}_k) = \text{mín} \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(x_{ik} - \bar{x}_k) \tag{1.25}$$

donde $d^2(x_{ik} - \bar{x}_k)$ es el cuadrado de la distancia euclidiana entre el elemento i del grupo k y su media de grupo. No es complicado probar que los criterios establecidos en (1.22) y (1.23) son equivalentes; nótese además que por ser $d^2(x_{ik} - \bar{x}_k)$ un escalar entonces trivialmente se tiene que $traza\,[d^2(x_{ik} - \bar{x}_k)] = d^2(x_{ik} - \bar{x}_k)$, de tal modo que

$$\text{mín} \sum_{k=1}^{K} \sum_{i=1}^{n_k} traza\left[d^2(x_{ik} - \bar{x}_k)\right] = \text{mín}\, traza \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^T (x_{ik} - \bar{x}_k)$$

y si llamamos W a la matriz de la suma de cuadrados dentro de los grupos, esto es

$$W = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^T (x_{ik} - \bar{x}_k)$$

se tiene entonces que

$$\text{mín}\, traza(W) = \text{mín}\, SCDG$$

Como la traza es la suma de los elementos de la diagonal principal ambos criterios coinciden. El algoritmo de K-medias busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro. El algoritmo funciona como sigue

1. Partir de una asignación inicial.
2. Comprobar si moviendo algún elemento se reduce W.
3. Si es posible reducir W mover el elemento, recalcular las medias de los dos grupos afectados por el cambio y volver a (2). Si no es posible reducir W terminar.

En consecuencia, el resultado del algoritmo puede depender de la asignación inicial y del orden de los elementos. Conviene siempre repetir el algoritmo desde distintos valores iniciales y permutando los elemento de la muestra.

Número de grupos

En la aplicación habitual del algoritmo de K-medias hay que fijar el número de grupos, K. Es claro que este número no puede estimarse con un criterio de homogeneidad ya que la forma de conseguir grupos muy homogéneos y minimizar la $SCDG$ es hacer tantos grupos como observaciones, con lo que siempre $SCDG = 0$. Se han propuesto distintos métodos para seleccionar el número de grupos. Un procedimiento aproximado que se utiliza bastante, aunque puede no estar justificado en unos datos concretos, es realizar un test F aproximado de reducción de variabilidad, comparando la$SCDG$ con K grupos con la de $K + 1$, y calculando la reducción proporcional de variabilidad que se obtiene aumentando un grupo adicional. El test es:

$$F = \frac{SCDG(K) - SCDG(K+1)}{SCDG(K+1)/(n - K - 1)}$$

y compara la disminución de variabilidad al aumentar un grupo con la varianza promedio.

Capítulo 2
Tarificación

En el mercado asegurador el precio de un servicio será aquel que permita cubrir los costos y provea un margen de utilidad a quien lo ofrece y sea aceptado por quien lo demanda.

Algunos principios fundamentales que deben tenerse en cuenta a la hora de calcular las tarifas son los siguientes:

1. Las tarifas deben ser suficientes para cubrir los costos de las reclamaciones más los gastos y proveer un margen de utilidad.

2. Las tarifas deben estar directamente relacionadas con el riesgo, esto es a mayor riesgo mayor tarifa.

3. Las tarifas deben ser el producto de la utilización de información estadística que cumpla exigencias de homogeneidad y representatividad.

2.1. Conceptos básicos

Para entender más a fondo el proceso de tarificación incluiremos algunos conceptos básicos asociados a este proceso [16].

Seguro
Actividad de servicios financieros por la que alguien se obliga mediante el cobro de una prima y para el caso que se produzca un evento determinado, a indemnizar a otro el daño producido.

Daño
Es la pérdida personal o material producida a consecuencia directa de un siniestro.

Prima
Es el precio pactado por el seguro contratado. Es la remuneración que recibe la aseguradora para hacerle frente a los riesgos que está amparando en la póliza y es la contraprestación que está obligando a ambas partes a cumplir con lo establecido

en el contrato.

Expuestos

Son el número de riesgos (pólizas) durante un período determinado, sobre los cuales en ese período la compañía debe dar cobertura y se determinan de la siguiente manera; Se toma un período determinado, para cada póliza se toma su vigencia y se cuenta cuántos períodos, meses, se encontraron activos dentro del período analizado, luego, para cada póliza, se divide el número de meses activos sobre el número de meses del período de análisis.

Indemnizaciones

Son los pagos que realizan las aseguradoras a los asegurados a consecuencia de pérdidas o daños a sus bienes o a sus personas. Las leyes de muchos países establecen que las indemnizaciones pueden ser en dinero o mediante la reposición de los bienes dañados por otros de las mismas características o condiciones. Esto es muy claro en el seguro de automóviles en donde la práctica es normalmente la reparación de los daños en los talleres con los que operan las aseguradoras y el asegurado no recibe ninguna cantidad de dinero por estos daños.

Siniestro

Es la materialización del riesgo. Quiere decir que es cuando sucede lo que se está amparando en la póliza y es motivo de indemnización.

Siniestralidad

Conjunto de siniestros producidos durante un periodo de tiempo determinado en un póliza o grupo de ellas. Por regla general se realiza en base a una cartera y en periodos anuales. También se puede entender este concepto como la proporción entre el importe total de las primas recaudadas por la entidad y el importe total de los siniestros.

Reclamaciones

Una reclamación es la solicitud de pago o reparación que hace un asegurado o beneficiario tras la ocurrencia de un siniestro.

Valor Asegurado

El valor asegurado corresponde al monto máximo pagadero en caso de siniestro, previamente estipulado en las condiciones de la póliza o sus anexos.

Gastos

Monto dinerario en que incurren la compañia para el ajuste y la liquidación de siniestros, así como para la emisión y mantenimiento de las pólizas. Usualmente suele usarse información histórica para su estimación.

Gastos de Ajuste de Siniestros

Los gastos de ajuste de siniestros son aquellos en los que debe incurrir la compañía para el ajuste y la liquidación de los siniestros (Costos del departamento de

indemnizaciones, honorarios jurídicos, etc.). Pueden dividirse en gastos asignables (se pueden asignar directamente a un siniestro) o gastos no asignables.

Frecuencia

La frecuencia (que en algunas ocasiones llamaremos frecuencia estadística) es una medida de la tasa de reclamaciones y usualmente se define como:

$$Frecuencia = \frac{No.\ de\ siniestros}{expuestos} \tag{2.1}$$

El análisis de los cambios en la frecuencia de las reclamaciones puede servir para identificar tendencias asociadas con la utilización de los servicios, efecto de las políticas de suscripción, entre otras cosas.

Severidad

La severidad (que en algunas ocasiones llamaremos severidad estadística) es el costo promedio por reclamación y se define como:

$$Severidad = \frac{Costos\ de\ los\ siniestros}{No.\ de\ siniestros} \tag{2.2}$$

El análisis de los cambios en la severidad provee información sobre el comportamiento de las pérdidas y sobre el impacto de los cambios en el manejo de las reclamaciones.

2.2. Agrupamiento de los datos

El cálculo de primas adecuadas depende en gran medida de la cantidad y calidad de la información que se emplea. La obtención y depuración de la información es parte fundamental dentro del proceso de tarifación.

Un paso fundamental en este proceso es el agrupamiento de los datos y dado que usualmente no existe la información suficiente para tarifar los riesgos a nivel individual, las compañías utilizan la tarificación por clases, que consiste en agrupar riesgos con pérdidas potenciales similares y estimar tasas diferentes para cada clase; para lograr esto es necesario establecer qué variables los segmentan de manera efectiva en grupos con pérdidas potenciales similares, para establecer tarifas equitativas. Las variables que se utilizan son llamadas variables de tarificación; los diferentes valores o rangos de estas variables se conocen como categorías o niveles de la variable y deben reflejar la variación en los costos esperados entre diferentes grupos de asegurados. Al realizar este tipo de agrupación se estaria evitando lo que se conoce como antiselección o selección adversa, que consiste en cobrar una misma tarifa a asegurados con diferente exposición de riesgo.

Dado el objetivo de equidad y suficiencia en las primas, es decir, contar con una prima igual para un grupo de asegurados con riesgos similares y así tener una prima suficiente para cubrir los posibles siniestros ocurridos, buscaremos la formación de

grupos de riesgo homogéneos determinados por combinaciones de clases de tarifa, que tendrán internamente una siniestralidad esperada similar y por lo tanto poca dispersión entorno a su valor esperado.

Para la generación de estas clases es indispensable emplear herramientas que permitan generar grupos con la suficiente robustez estadística para generar estimadores más confiables. Dentro de las diversas técnicas empleadas con este fin, se optó por usar el método Conglomerados (Clusters) para agrupar las observaciones en conjuntos con riesgos similares, ya que además de ser fácil de implementar tiene una gran variedad de métodos para agrupar como los vistos en la **Sección 1.3** lo cual nos permitirá tener más elementos para comparar y así seleccionar los que mejor se adecuen al propósito de este trabajo.

2.3. Método de tarificación

El propósito de la tarifación consiste en asegurar que las primas sean suficientes para cubrir todos los costos asociados con la transferencia del riesgo y se obtenga un cierto margen de utilidad, sin que las primas sean excesivas. Para lograr dicho objetivo, es necesario que la tarifación sea prospectiva, esto es, que refleje las condiciones del momento en que las tarifas se van a aplicar.

Uno de los métodos más empleados en el cálculo de las primas es el método de prima pura de riesgo, para utilizar este método es necesario estimar y proyectar las pérdidas del período de experiencia (período observado), que son indicativas de las pérdidas en el período de proyección (período en el cual las tarifas estarán vigentes).

2.3.1. Prima Pura

La prima pura o también conocida como prima de riesgo corresponde al costo promedio de todos los siniestros por expuestos, esto es:

$$\begin{aligned} Prima\ Pura &= \frac{Costos\ de\ los\ siniestros}{expuestos} \\ &= \frac{No.\ de\ siniestros}{expuestos}\ \frac{Costos\ de\ los\ siniestro}{No.\ de\ siniestros} \\ &= Frecuencia * Severidad \end{aligned} \tag{2.3}$$

El término prima pura se refiere básicamente a los costos esperados, que corresponden únicamente a los siniestros. No se incluyen gastos ni margen de utilidad. Usualmente la Prima Pura es cargada con un factor de gastos y utilidad, para obtener la prima comercial.

En este método de prima pura, se estiman las pérdidas totales o pérdidas últimas y se calcula el número de expuestos correspondiente al período de experiencia.

Posteriormente las pérdidas totales son proyectadas teniendo en cuenta aspectos tales como la inflación, para que sean representativas de las que se observarán en el período en el cual las tarifas estarán vigentes.

Dado que la tarifación utilizando este método puede presentar distorsiones, debido principalmente a que no siempre se tienen en cuenta las relaciones entre las diferentes variables de tarifación y por consiguiente generan una gran volatilidad, en los últimos años se ha extendido el uso de métodos multivariados, tales como los modelos lineales generalizados, con el objetivo de solventar estos problemas. El uso de métodos multivariados permite modelar todas las variables a la vez y esto sirve para reducir la volatilidad que presenta el método anterior. También permite asumir una distribución para la frecuencia y para la severidad, establecer mediante criterios estadísticos cuáles son las variables más relevantes, los cuales sirven para evaluar qué tan apropiado es el modelo que se está asumiendo, entre otras ventajas.

2.3.2. Prima Pura GLM

Actualmente los modelos lineales generalizados son el modelo estándar de tarifación por clases ya que permiten calcular el valor esperado de la variable respuesta dado un conjunto de variables explicativas. Las variables de respuesta no tienen que estar distribuidas normalmente, pero deben pertenecer a la familia exponencial. Así mismo, la relación entre el valor esperado de la variable respuesta y las variables explicativas no tiene que ser necesariamente lineal, lo que se pide es que una función de la media, conocida como función de enlace (o función liga), sea lineal con respecto a las variables explicativas. El esquema principal para tarifar seguros usando este método consiste en modelar la frecuencia y la severidad por separado y multiplicar los resultados obtenidos para obtener la prima pura.

Como se vió en el primer capítulo en un modelo lineal la variable dependiente Y se distribuye normalmente y su valor esperado $\mathbb{E}(y_i) = \mu_i$ para la observación $i-$ésima $(i = 1, 2, ..., n)$, se modela directamente como una función de $(p-1)$ variables independientes $x_1, x_2, ..., x_{p-1}$ a través de un predictor lineal η_i de la siguiente forma:

$$\mathbb{E}(y_i) = \mu_i = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} = \sum_j \beta_j x_{ij} \tag{2.4}$$

En los Modelos Lineales Generalizados se permite que la variable dependiente Y siga cualquier distribución que pertenezca a la familia exponencial, además en vez de modelar el valor esperado de la variable dependiente como en (2.4) lo podemos modelar a través de una función liga $g(\cdot)$ como:

$$g(\mathbb{E}(y_i)) = g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} \tag{2.5}$$

Para obtener la prima de riesgo lo que se busca es modelar la frecuencia y severidad con algunas distribuciones que pertenezcan a la familia exponencial y una función liga es el logaritmo natural, así al usar la función inversa (la exponencial) el valor esperado de la variable dependiente estará dado por:

$$\mathbb{E}(y_i) = \mu_i = g^{-1}(\eta_i) = e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)}} \tag{2.6}$$

Frecuencia

Para modelar la frecuencia, se debe tener en cuenta que la variable dependiente es una variable de conteo donde el número de siniestros depende de la exposición (expuestos), así el valor esperado de esta variable queda de la siguiente manera:

$$g\left[\mathbb{E}(y_i)\right] = g\left[\mathbb{E}\left(\frac{ns_i}{\theta_i}\right)\right] = g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} \tag{2.7}$$

Donde ns_i son el número de siniestros, θ_i son los expuestos que son conocidos y g es la liga logarítmo natural lo cual implica lo siguiente:

$$\ln \mathbb{E}\left(\frac{ns_i}{\theta_i}\right) = \ln \mu_i = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} \tag{2.8}$$

por tanto se tiene:

$$\begin{aligned} \ln \mathbb{E}\left(\frac{ns_i}{\theta_i}\right) &= \ln \mathbb{E}(ns_i) - \ln \theta_i \\ &= \ln \mu_i = \eta_i \\ &= \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} \end{aligned}$$

De esta manera se sigue lo siguiente:

$$\begin{aligned} \ln \mathbb{E}(ns_i) &= \ln \mu_i + \ln \theta_i \\ &= \ln(\mu_i * \theta_i) = \ln \mu_i^* \\ &= \ln \theta_i + \left(\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)}\right) \end{aligned}$$

Entonces al aplicarle la función exponencial se tiene

$$\begin{aligned} \mathbb{E}(ns_i) = \mu_i^* &= g^{-1}(\eta_i) \\ &= e^{\ln \theta_i + \left(\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)}\right)} \\ &= e^{\ln \theta_i} e^{\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)}\right)} \\ &= \theta_i \ e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)}} \end{aligned} \tag{2.9}$$

Al parametro θ_i se le conoce tambien como el offset.

Así entonces para modelar la frecuencia se tiene que la variable dependiente será el número de siniestros y la variable offset será el logaritmo de los expuestos.

Severidad

Para modelar la severidad se tiene que el valor esperado de la variable de respuesta queda de la siguiente manera:

$$g\left[\mathbb{E}(y_i)\right] = g\left[\mathbb{E}\left(\frac{cs_i}{ns_i}\right)\right] = g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} \tag{2.10}$$

Donde ns_i son el número de siniestros cs_i son los costos de los siniestros y g es la liga logarítmo natural, lo cual implica lo siguiente:

$$\ln \mathbb{E}\left(\frac{cs_i}{ns_i}\right) = \ln \mu_i = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} \tag{2.11}$$

Entonces al aplicarle la función exponencial se tiene

$$\mathbb{E}(y_i) = \mu_i^* = g^{-1}(\eta_i) = e^{\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)}\right)} \tag{2.12}$$

Así entonces para modelar la severidad se tiene que la variable dependiente sera el costo promedio del número de siniestros ocurridos (severidad).

Variables Categóricas

En el desarrollo que se ha realizado hasta el momento, se ha supuesto que las variables independientes son continuas, y en este caso, la relación descrita para modelar el valor esperado funciona adecuadamente. Sin embargo, cuando se tienen variables categóricas, se deben tomar en cuenta los niveles de estas variables, por tal motivo, es necesario formular la representación adecuada del valor esperado de la variable de respuesta.

Sean $V1_{n1}, V2_{n2}, ..., Vm_{nm}$ variables categóricas con nj $j = 1, 2, ..., m$ niveles de cada variable, así el valor esperado de la variable de respuesta queda de la siguiente manera:

$$\begin{aligned} \mathbb{E}(y_i) &= \mu_i = g^{-1}(\eta_i) \\ &= e^{(\beta_0 + \beta_{11} V1_1 + ... + \beta_{1n1} V1_{n1} + \beta_{21} V2_1 + ... + \beta_{2n2} V2_{n2} + ... + \beta_{m1} Vm_1 + ... + \beta_{mnm} Vm_{nm})} \\ &= e^{(\beta_0 + \sum_{n1} \beta_{1n1} V1_{n1} + ... + \sum_{nm} \beta_{mnm} Vm_{nm})} \end{aligned} \tag{2.13}$$

donde para $j = 1, 2, ..., m$ se tiene

$$Vj_{nj} = \begin{cases} 1 & \text{si} \quad y_i \in Vj_{nj} \\ 0 & \text{si} \quad y_i \notin Vj_{nj} \end{cases}$$

donde β_0 es el valor promedio, comúnmente recibe el nombre de intercepto .

Frecuencia

Sean Fj variables categóricas con nj niveles de cada variable respectivamente $j = 1, ..., p$, aplicando la relación (2.13) y de (2.9) se tiene que la frecuencia para la variable dependiente y_i queda de la siguiente manera:

$$Frecuencia(y_i) = \mu_i^F = g^{-1}(\eta_i^F) = e^{(\beta_0 + \sum_{n1} \beta_{1n1} F1_{n1} + ... + \sum_{np} \beta_{pnp} Fp_{np})} \tag{2.14}$$

donde $\beta_0 = \ln \theta_i + cte$ es el intercepto, θ_i son los expuestos y para $j = 1, ..., p$ se tiene que:

$$Fj_{nj} = \begin{cases} 1 & \text{si} \quad y_i \in Fj_{nj} \\ 0 & \text{si} \quad y_i \notin Fj_{nj} \end{cases}$$

en este caso la variable offset es $\ln \theta_i$.

Severidad

Sean Sk variables categóricas con mk niveles de cada variable respectivamente $k = 1, ..., q$, aplicando la relación (2.13) y de (2.12) se tiene que la severidad para la variable dependiente y_i queda de la siguiente manera:

$$Severidad(y_i) = \mu_i^S = g^{-1}(\eta_i^S) = e^{(\beta_0^* + \sum_{m1} \beta_{1m1}^* S1_{m1} + ... + \sum_{mq} \beta_{qmq}^* Sq_{mq})} \tag{2.15}$$

donde β_0^* es el intercepto y para $k = 1, ..., q$ se tiene que:

$$Sk_{mk} = \begin{cases} 1 & \text{si} \quad y_i \in Sk_{mk} \\ 0 & \text{si} \quad y_i \notin Sk_{mk} \end{cases}$$

Prima pura o Prima de Riesgo

Una vez definidas las variable que influyen en la modelación de la frecuencia y severidad y como se definió en la sección 2.3.1 la prima de riesgo es el producto de la frecuencia por la severidad, así por (2.14) y (2.15) se tiene que la Prima pura de riesgo para la variable y_i queda de la siguiente manera:

$$\begin{aligned} Prima(y_i) &= Frecuencia * Severidad \\ &= e^{(\beta_0 + \sum_{n1} \beta_{1n1} F1_{n1} + ... + \sum_{np} \beta_{pnp} Fp_{np})} * e^{(\beta_0^* + \sum_{m1} \beta_{1m1}^* S1_{m1} + ... + \sum_{mq} \beta_{qmq}^* Sq_{mq})} \end{aligned} \tag{2.16}$$

El cálculo de los parámetros involucrados en las regresiones se realiza a través de estimaciones de máxima verosimilitud y dada la complejidad de las funciones de verosimilitud, no hay métodos analíticos para hallar los puntos máximos de dichas funciones, por tal motivo es necesario utilizar un software que estime dichos parámetros.

2.4. Distribuciones adecuadas

Como se ha visto hasta el momento, el cálculo de la prima de riesgo para una cobertura en seguros de automóviles, está basado en la adecuada modelación de la frecuencia y de la severidad, por lo que una parte fundamental en dicha modelación es considerar la distribución que mejor represente el comportamiento de los datos observados.

Así teniendo en cuenta estas consideraciones, se presentan a continuación las distribuciones de probabilidad más utilizadas para modelar cada una de estas variables aleatorias.

2.4.1. Distribuciones para la Frecuencia

La frecuencia se modelará utilizando distribuciones de probabilidad discretas de valores no negativos, se utilizan estas distribuciones debido a que permiten determinar la ocurrencia o no ocurrencia de un evento con algún tipo de riesgo, en el campo actuarial las distribuciones más utilizadas son la Poisson y la Binomial Negativa.

Distribución Poisson

El modelo Poisson asume que todos los asegurados tienen el mismo riesgo subyacente, es decir, la ocurrencia de un siniestro constituye un evento aleatorio y no hay razón para penalizar a los asegurados responsables por aquel siniestro; es decir dado el número de siniestros en un intervalo de tiempo se tienen que formular tres supuestos, los cuales darán origen a la función de probabilidad de que ocurran cierto número de siniestros:

1. Mientras más tiempo una persona conduce mayor probabilidad tiene de un siniestro.

2. La probabilidad de que ocurra más de un siniestro es muy pequeña.

3. Si tenemos dos intervalos de tiempo separados entonces el número de siniestros relacionados con esos intervalos de tiempo son independientes.

Por consiguiente se tendra una distribución que expresa a partir de una frecuencia de ocurrencia media λ, la probabilidad de que ocurra un determinado número de eventos durante un cierto periodo de tiempo.

Distribución Binomial Negativa

En este modelo suponemos que los asegurados no tienen el mismo riesgo subyacente, es decir el comportamiento de los asegurados es heterogéneo ya que el riesgo de tener un siniestro es significativamente diferente entre cada asegurado y por tanto se justifica el recargo o el descuento en la prima a pagar ya que se refleja directamente en función de su riesgo.

2.4.2. Distribuciones para la Severidad

Para este caso, la severidad se modelará ajustando distribuciones de probabilidad de carácter continuo, por tanto se buscan modelos de distribución probabilística que se ajusten a una serie de datos históricos de pérdidas.

Otras distribuciones que pueden ser consideradas son la distribución Lognormal y la distribución Weibull (pero no se descartan otras opciones como Gamma, Beta, Pareto), aunque en la práctica ninguna distribución simple se ajusta a los datos satisfactoriamente.

Distribución log-Normal

La distribución log-normal es usada habitualmente para modelar siniestros de cola larga, es decir, donde se den valores muéstrales muy alejados de la media o coeficientes de asimetría positivos muy elevados.

Una distribución log-normal es una distribución de probabilidad de una variable aleatoria cuyo logaritmo está normalmente distribuido, o lo que es lo mismo si ξ es una variable aleatoria que sigue una distribución normal, entonces e^{ξ} sigue una log-normal.

Distribución Gamma

La distribución gamma es muy utilizada habitualmente en la práctica para modelar siniestros de cola corta o exponencial, es decir, aquellos que no presentan valores muéstrales muy alejados de la media. Esta distribución suele ajustarse bien a riesgos de carácter no catastrófico.

2.5. Selección del Modelo

Para seleccionar cuáles son las variables que mejor explican a la variable respuesta, de acuerdo con la calidad del ajuste y el número de parámetros necesitamos tener en cuenta algunos criterios estadísticos para elegir un mejor modelo.

2.5.1. Estimadores

Una vez estimado el modelo, partiendo de que se ha realizado un muestreo probabilístico, nos interesa contrastar si los coeficientes estimados son significativamente distintos de 0. Es decir, si una determinada variable explicativa tiene un efecto significativo sobre la respuesta o no. Para dar respuesta a esta pregunta, utilizamos los contrastes de hipótesis sobre los parámetros, explicando un método para realizarlo, el contraste basado en la prueba de Wald.

2.5.2. Contraste de Wald

Este contraste está basado en la normalidad asintótica de los estimadores. Se requiere contrastar si un parámetro $\beta_r = 0$, frente a que no lo sea, es decir:

$$H_0 : \beta_r = 0$$

$$vs$$

$$H_1 : \beta_r \neq 0$$

Wald, demostró que bajo la hipótesis nula

$$\frac{\widehat{\beta_r}}{\widehat{SE(\beta_r)}} \rightarrow N(0,1)$$

donde, $\widehat{\beta_r}$ y $\widehat{SE(\beta_r)}$ son las estimaciones del modelo para β_r y el error estándar de β_r, de esta forma el estadístico de contraste es:

$$W_r = \frac{\widehat{\beta_r}}{\widehat{SE(\beta_r)}}$$

Intervalos de Confianza de los parámetros

Estos intervalos se basan en que los parámetros β_r siguen asintóticamente una distribución $N(\beta_r, \hat{\sigma}^2(\widehat{\beta_r}))$ con lo que

$$P\left[-z_{\frac{\alpha}{2}} \leq \frac{\widehat{\beta_r} - \beta_r}{\widehat{\sigma}(\widehat{\beta_r})} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

Así el intervalo de confianza para β_r a un nivel $(1 - \alpha)$ está dado por:

$$\widehat{\beta_r} \pm z_{\frac{\alpha}{2}} \widehat{\sigma}(\widehat{\beta_r})$$

Si el intervalo de confianza incluye al cero, significa que al nivel α elegido no se podría rechazar la hipótesis nula de que $\beta_r = 0$

2.5.3. Medidas de Bondad de Ajuste

El objetivo de la evaluación de modelos es seleccionar aquellos que presenten el mejor balance entre la capacidad de ajuste de los datos y su complejidad. Modelos con un mayor número de parámetros tienden a ajustar mejor una base de datos, no obstante tienden a ser más inestables y a modelar la variabilidad de esos datos más que su tendencia.

Cuando se tienen dos o más modelos, una metodología para compararlos es a través de la la función de máxima verosimilitud, que nos permite seleccionar el modelo que mejor se ajusta a los datos observados; con base en esto existen medidas de contraste que hacen uso de la log verosimilitud l y sustraen un término proporcional al número de parámetros p en el modelo, entre los más conocidos están, El Criterio de Información Akaike (AIC) y el Criterio de Información Bayesiano (BIC).

Criterio de Información de Akaike (AIC)

Este criterio tiene en cuenta los cambios en la bondad de ajuste y las diferencias en el número de parámetros entre dos modelos. Los mejores modelos son aquellos que presentaron el menor valor de AIC y se define como:

$$AIC = -2l + 2p$$

Cuando los valores de AIC están muy cercanos la selección del mejor modelo se puede realizar con base en el cálculo de la probabilidad (pesos de Akaike) y la probabilidad relativa (relación de evidencia), a través de las siguientes ecuaciones:

$$Probabilidad = \frac{e^{-0,5^{\Delta}}}{1 + e^{-0,5^{\Delta}}}$$

$$Probabilidad\ \ Relativa = \frac{Prob\ \ de\ \ que\ \ el\ \ modelo\ \ 1\ \ sea\ \ correcto}{Prob\ \ de\ \ que\ \ el\ \ modelo\ \ 2\ \ sea\ \ correcto} = \frac{1}{e^{-0,5^{\Delta}}}$$

Donde *Delta* es la diferencia entre los valores de AIC.

Criterio de Información Bayesiano (BIC)

Este criterio es calculado para los diferentes modelos como una función de bondad de ajuste de la logverosimilitud l el número de parámetros ajustados p y el número total de datos n. El modelo con el más bajo valor de BIC es considerado el que mejor explica los datos con un mínimo número de parámetros. El BIC está definido de la siguiente manera:

$$BIC = -2l + plog(n)$$

Los modelos comparados utilizando AIC o BIC deben estar basadas en el mismo conjunto de observaciones. Esto afecta a la hora de decidir entre los modelos que incluyen al menos una variable explicativa con valores perdidos.

Un modelo con un gran número de parámetros y un buen ajuste tiene un valor bajo para el término de sesgo, pero el criterio es empujado hacia arriba por el término varianza. Por el contrario un modelo con p pequeño tiene un mayor término de sesgo, pero menor término varianza. La decisión tiene que ser hecho si usar AIC o BIC. Esto último se aplica una mayor pena por el número de parámetros, por lo que tiende a elegir modelos con un menor número de variables explicativas en comparación con AIC. Cuando n es grande, como es el caso en la mayoría de los conjuntos de datos seguros, el BIC tiende a selecciona modelo que la mayoría de los analistas considere demasiado simple. En este caso, la AIC es preferible.

2.6. Intervalos de confianza

Una vez que se ya se hayan seleccionados los modelos adecuados para la frecuencia y la severidad y así poder calcular la prima de riesgo, es necesario dar un intervalo de confianza para nuestra estimación. Los intervalos de confianza que se proponen obtener, son sobre el valor estimado $\hat{\mu}$ de la media de la respuesta $\mathbb{E}(y) = \mu$.

Antes de continuar con los intervalos de confianza es necesario representar la ecuación (1.2) de una manera más general como:

$$f(y;\theta) = c(y,\phi)\exp\left(\frac{y\theta - a(\theta)}{\phi}\right) \tag{2.17}$$

donde ϕ es el parámetro de dispersión y

$$a''(\theta) = \frac{\partial a'(\theta)}{\partial \theta} = \frac{\partial \mu}{\partial \theta} \equiv V(\mu)$$

Con esta representación de la familia exponencial se tiene que la media de la frecuencia μ_i^F y la media de la severidad μ_i^S se modelan con liga logarítmica, por lo que la estimación de ambas cantidades se forma al sustituir los valores de los estimadores de los parámetros para la frecuencia $\hat{\beta}$ y para la severidad $\hat{\beta}^*$ con sus respectivas covariables categóricas $\mathbf{x}_i^F = \mathbf{F}_i$ y $\mathbf{x}_i^S = \mathbf{S}_i$:

$$\mu_i^F = \exp((\mathbf{x}_i^F)^T \widehat{\beta}) = \exp(\mathbf{F}_i^T \widehat{\beta})$$

$$\mu_i^S = \exp((\mathbf{x}_i^S)^T \widehat{\beta^*}) = \exp(\mathbf{S}_i^T \widehat{\beta^*})$$

Para hallar el intervalo de confianza de $\widehat{\mu_i}$ hay que calcular primero la varianza del predictor lineal $\mathbf{x}_i^T \beta$. Se tiene que:

$$Var(\mathbf{x}_i^T \widehat{\beta}) = \phi \mathbf{x}_i^T (X^T W X)^{-1} \mathbf{x}_i$$

donde X es la matriz formada por la información de todas las observaciones $\mathbf{x}_i$ y W es una matriz diagonal con entradas $[\{\dot{g}(\mu_i)\}^2 V(\mu_i)]^{-1}$.

La varianza del predictor lineal queda de esta forma debido a que la estimación de los parámetros se hace empleando máxima verosimilitud, por lo que el vector de parámetros tiene una distribución asintótica normal [12]:

$$\widehat{\beta} \sim N(\beta, \phi(X^T W X)^{-1})$$

Una vez estimada la varianza del componente lineal, se puede hacer un intervalo de confianza para la componente lineal $\mathbf{x}_i^T \beta$ como $(g(\mu_l), g(\mu_u))$:

$$g(\mu_l) = \mathbf{x}_i^T \widehat{\beta} - z_{1-\alpha} \sqrt{\widehat{\phi} \mathbf{x}_i^T (X^T \widehat{W} X)^{-1}} \mathbf{x}_i$$

$$g(\mu_u) = \mathbf{x}_i^T \widehat{\beta} + z_{1-\alpha} \sqrt{\widehat{\phi} \mathbf{x}_i^T (X^T \widehat{W} X)^{-1}} \mathbf{x}_i$$

Finalmente se aplica la función inversa de la función liga para obtener un intervalo de confianza para (μ_l, μ_u).

Con el procedimiento anterior se pueden calcular intervalos de confianza y por consiguiente la varianza de $\widehat{\mu}_i^F$ y $\widehat{\mu}_i^S$, por tanto para tener un intervalo de la prima de riesgo estimada podemos utilizar el método Delta para calcular el error estándar de $(\widehat{\mu}_i^F * \widehat{\mu}_i^S)$. El método delta establece que una aproximación de la varianza de una función $g(t)$ dada por:

$$Var(g(\mathbf{t})) \approx \sum_{i=1}^{k} g_i'(\theta)^2 Var(t_i) + 2 \sum_{i>j} g_i'(\theta) g_j'(\theta) Cov(t_i, t_j) \tag{2.18}$$

la aproximación del valor esperado de $g(t)$ esta dado por:

$$\mathbb{E}(g(t)) \approx g(\theta)$$

Así, el valor esperado es simplemente la función, donde $g(t)$ es: $g(\widehat{\mu}_i^F, \widehat{\mu}_i^S)$. El valor esperado de $g(\widehat{\mu}_i^F, \widehat{\mu}_i^S) = \widehat{\mu}_i^F \widehat{\mu}_i^S$, para la varianza necesitamos las derivadas parciales de $g(\widehat{\mu}_i^F, \widehat{\mu}_i^S)$:

$$\frac{\partial}{\partial \widehat{\mu}_i^F} g(\widehat{\mu}_i^F, \widehat{\mu}_i^S) = \widehat{\mu}_i^S$$

$$\frac{\partial}{\partial \widehat{\mu}_i^S} g(\widehat{\mu}_i^F, \widehat{\mu}_i^S) = \widehat{\mu}_i^F$$

De (2.18) tenemos

$$Var(\widehat{\mu}_i^F \widehat{\mu}_i^S) = (\widehat{\mu}_i^S)^2 Var(\widehat{\mu}_i^F) + (\widehat{\mu}_i^F)^2 Var(\widehat{\mu}_i^S) + 2\widehat{\mu}_i^F \widehat{\mu}_i^S Cov(\widehat{\mu}_i^F, \widehat{\mu}_i^S) \tag{2.19}$$

El error estándar sería simplemente la raíz cuadrada de (2.19), así el intervalo de confianza a un 95 % para $\widehat{\mu}_i^F \widehat{\mu}_i^S$ estará dado por:

$$\widehat{\mu}_i^F \widehat{\mu}_i^S \pm 1{,}96 \widehat{SE}(\widehat{\mu}_i^F \widehat{\mu}_i^S) \tag{2.20}$$

Capítulo 3
Aplicación

Una vez descrito los conceptos estadísticos y teóricos sobre la estimación de la prima pura de riesgo en seguros de automóviles, es el turno de emplear estas herramientas para generar una estimación adecuada de las mismas, que permita a las aseguradoras cubrir las reclamaciones de los sinestros que los asegurados tengan durante la vigencia de su póliza.

Como se mencionó en la introducción de este proyecto, se realizará el cálculo de la prima de riesgo con información de todo el sector asegurador. Se cuenta con una base de datos a nivel nacional[1] con las variables necesarias para el cálculo de la prima de riesgo (Número de Siniestros, Unidades Expuestas, Siniestro Ocurrido) que resumen el comportamiento sectorial del año 2014[2]. Por fines prácticos e ilustrativos, se trabajará solamente con las carrocerías de la marca TOYOTA, aplicando el método descrito en el capítulo anterior para las cuatro principales coberturas: daños materiales, robo total, gastos médicos y responsabilidad civil.

3.1. Agrupamiento de datos

Como se vio en la **Sección 2.3.2** nuestro modelo contará con variables categóricas, las cuales serán el Estado de circulación del Automóvil y tipo de Carrocería, por lo tanto, para tener los niveles de estas variables será necesario agrupar los 32 Estados y las 18 carrocerías en grupos lo mas homogéneos posibles con el propósito de obtener una tarifa justa a cobrar al asegurado y que sea suficiente para la contraparte.

Desglosando la base de datos proporcionada por AMIS, se calculará la frecuencia y severidad por cobertura para los estados y carrocerías por separado, esto es con el propósito de formar el número de grupos necesarios para ser considerados como niveles de las variables categóricas; así por lo visto en la **Sección 1.3** del **Capítulo**

[1]Fuente AMIS (Asociación Mexicana de Instituciones de Seguros)

[2]La información solicitada fue un requerimiento especial y con fines solo académicos, no lucrativos; donde se evaluó disponer de una base muestra de algunas marcas y bajo los lineamientos donde no es posible identificar información de las compañías y no se vulneran las políticas de confidencialidad.

1 podemos aplicar algún método de clusterización para obtener el número de grupos necesarios.

En las siguientes dos subsecciones, se mostrarán las técnicas de agrupamiento para la frecuencia y severidad de la cobertura de Daños Materiales; los grupos para la frecuencia y la severidad de las demás coberturas se muestran en la **Sección 3.1.3.**

3.1.1. Frecuencia de Daños Materiales

A continuación se formarán los grupos por estado y carrocería teniendo en cuenta su frecuencia estadística que nos servirán para modelar mediante GLM's la frecuencia con la cual estimaremos la prima pura de riesgo.

Estados

Para la siguiente agrupación de Estados se utilizó la liga Ward, teniendo como resultado lo siguiente:

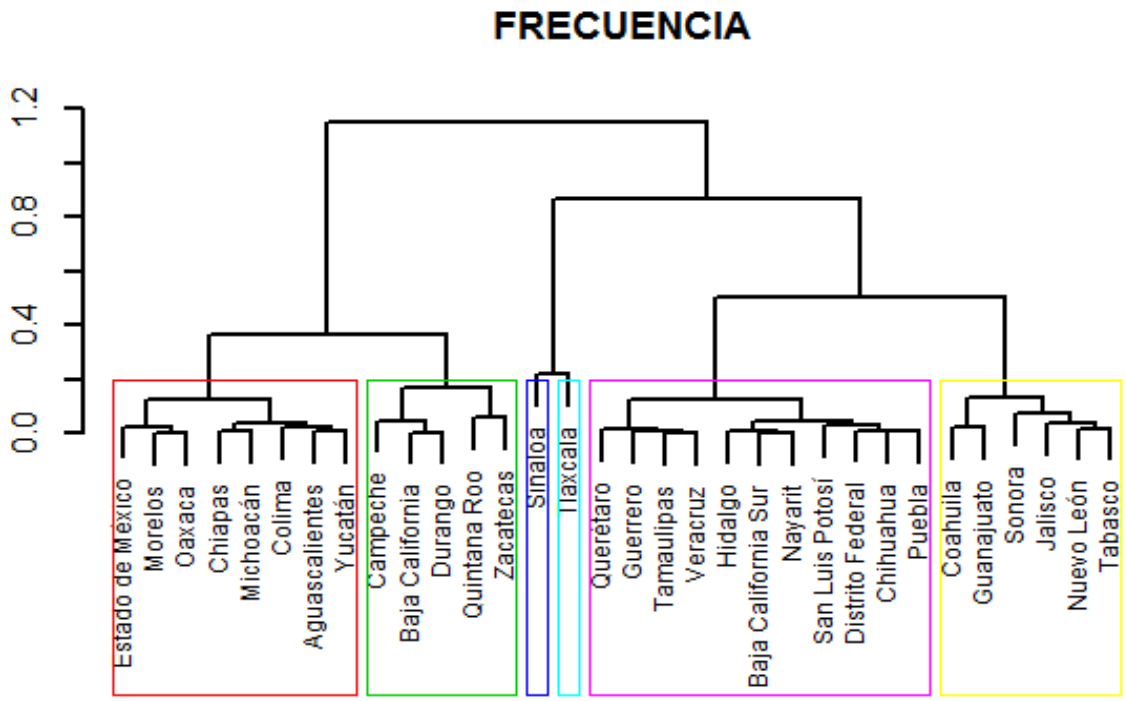


Figura 3.1: Agrupación de estados para la frecuencia estadística de la cobertura de Daños Materiales

Como se observa en la Figura 3.1 se tomarán 6 grupos, donde los estados de Sinaloa y Tlaxcala formarán un grupo cada uno ya que estos estados son los que presentan una mayor frecuencia de números de siniestros y los cuatros grupos restantes se conforman de estados con frecuencias similares.

Carrocerías

Para el caso de las carrocerías se utilizará la liga simple, teniendo como resultado la siguiente agrupación:

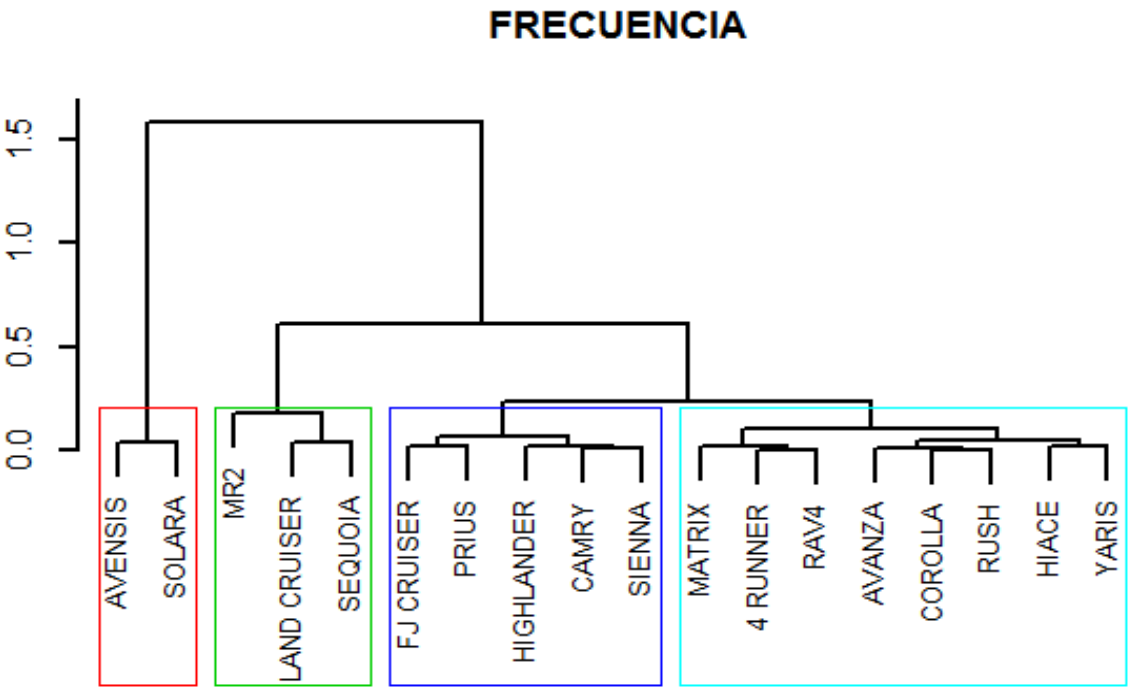


Figura 3.2: Agrupación de Carrocerías para la frecuencia estadística de la cobertura de Daños Materiales

En este caso, de la Figura 3.2 se tienen cuatro grupos de carrocerías con frecuencia de número de siniestros similares.

3.1.2. Severidad de Daños Materiales

Al igual que con la frecuencia, aplicaremos un método de clusterización para agrupar los estados y carrocerías por severidades estadísticas de la cobertura de Daños Materiales similares para posteriormente estimar la severidad que junto con la frecuencia estimada nos ayudará a modelar la prima pura de riesgo.

Estados

Para agrupar los estados por severidad utilizaremos la liga Ward, teniendo como resultado la siguiente agrupación:

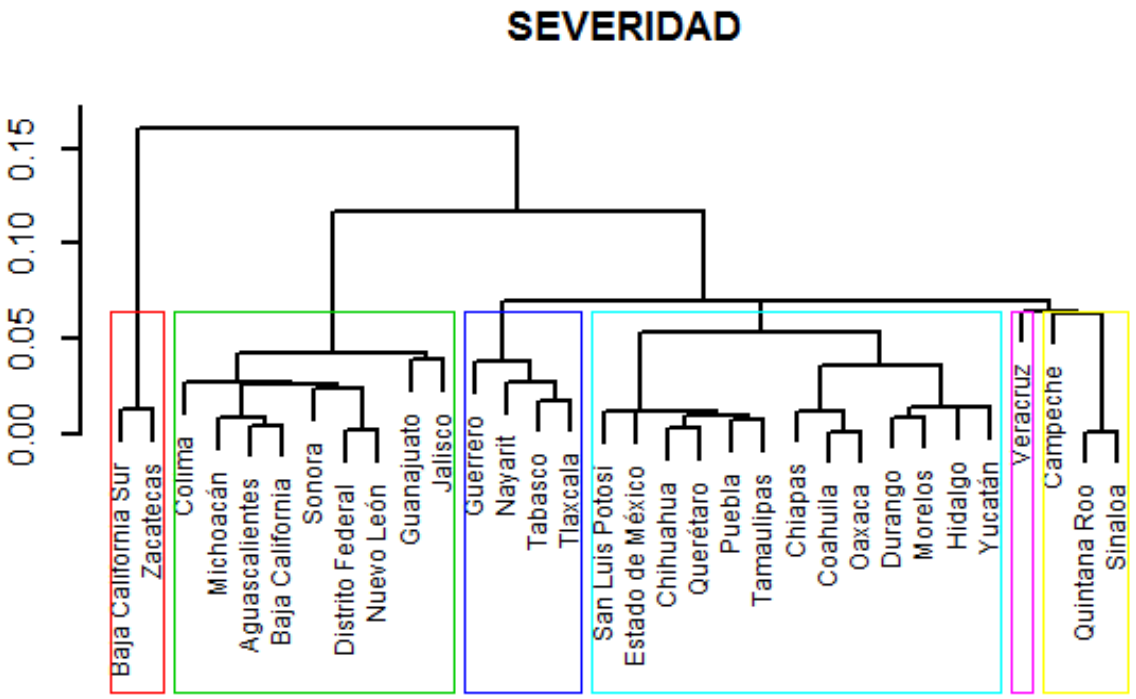


Figura 3.3: Agrupación de Estados para la severidad estadística de la cobertura de Daños Materiales

Como se observa en la Figura 3.3 se tomarán de nuevo 6 grupos, en este caso, el estado que formará un grupo es Veracruz ya que tiene una mayor severidad reportada que el resto de los estados, los grupos restantes se conforman de estados con comportamiento de severidades similares.

Carrocería

Para agrupar las carrocerías se utilizará la liga simple, teniendo como resultado la siguiente agrupación:

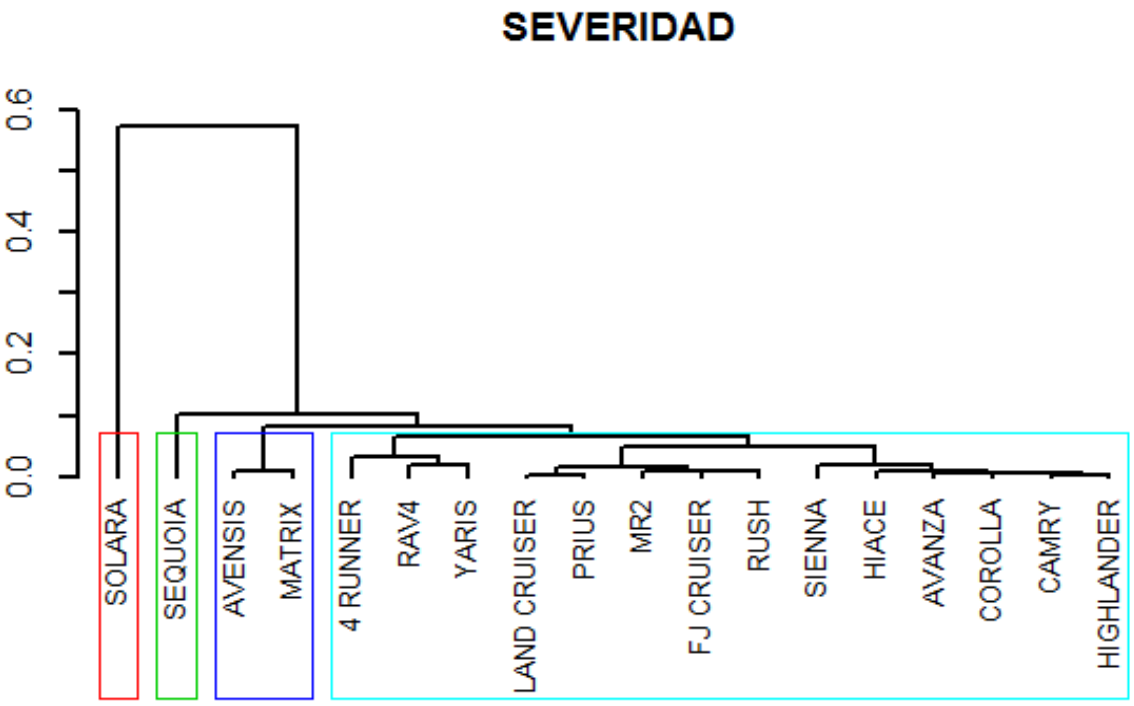


Figura 3.4: Agrupación de Carrocerías para la severidad estadística de la cobertura de Daños Materiales

De la Figura 3.4 , se tienen que las carrocerías SOLARA Y SEQUOIA forman un grupo cada una, esto debido a que sus severidades estadísticas son mayores que el resto de las demás carrocerías, teniendo un grupo formado por 14 carrocerías con severidades estadísticas muy similares.

3.1.3. Grupos

Una vez que ya se detectaron los estados y carrocerías con frecuencia y severidades estadísticas similares, resumiremos esos resultados en una tabla etiquetando la carrocería y estado según el grupo asignado[3].

En la siguiente figura se muestra una tabla con las agrupaciones por estados según frecuencia o severidad de las cuatro coberturas:

[3]Los grupos presentados son los que mejor se ajustabán a los modelos, previamente se probaron utilizando diferentes números de grupos y diferentes ligas.

Agrupación de Estados por Cobertura								
ESTADO	Daños Materiales		Robo Total		Responsabilidad Civil		Gastos Médicos	
	Frecuencia	Severidad	Frecuencia	Severidad	Frecuencia	Severidad	Frecuencia	Severidad
Aguascalientes	E1	E1	E1	E1	E1	E1	E1	E1
Baja California	E2	E1	E2	E2	E2	E1	E1	E2
Baja California Sur	E3	E2	E1	E1	E2	E1	E2	E3
Campeche	E2	E3	E1	E1	E3	E1	E2	E2
Chiapas	E1	E4	E1	E3	E1	E1	E1	E2
Chihuahua	E3	E4	E3	E3	E3	E1	E3	E2
Coahuila	E4	E4	E3	E3	E3	E1	E3	E1
Colima	E1	E1	E1	E4	E4	E2	E1	E1
Distrito Federal	E3	E1	E3	E5	E5	E1	E1	E2
Durango	E2	E4	E3	E4	E4	E1	E3	E2
Estado de México	E1	E4	E4	E5	E3	E1	E1	E2
Guanajuato	E4	E1	E1	E2	E4	E1	E1	E1
Guerrero	E3	E5	E5	E3	E1	E2	E1	E3
Hidalgo	E3	E4	E1	E3	E4	E2	E1	E1
Jalisco	E4	E1	E2	E5	E5	E1	E3	E2
Michoacán	E1	E1	E3	E4	E2	E2	E1	E1
Morelos	E1	E4	E2	E3	E3	E1	E1	E1
Nayarit	E3	E5	E1	E1	E4	E2	E3	E3
Nuevo León	E4	E1	E1	E5	E5	E1	E1	E2
Oaxaca	E1	E4	E3	E3	E2	E2	E2	E3
Puebla	E3	E4	E3	E5	E3	E1	E3	E2
Querétaro	E3	E4	E1	E4	E1	E1	E1	E1
Quintana Roo	E2	E3	E1	E2	E3	E1	E1	E3
San Luis Potosí	E3	E4	E1	E2	E3	E1	E3	E1
Sinaloa	E5	E3	E6	E4	E5	E2	E3	E1
Sonora	E4	E1	E3	E3	E1	E1	E3	E1
Tabasco	E4	E5	E3	E3	E3	E2	E1	E1
Tamaulipas	E3	E4	E2	E5	E1	E2	E1	E1
Tlaxcala	E6	E5	E3	E2	E4	E2	E4	E2
Veracruz	E3	E6	E3	E5	E1	E2	E1	E1
Yucatán	E1	E4	E1	E2	E5	E2	E3	E2
Zacatecas	E2	E2	E2	E3	E4	E2	E2	E3

Figura 3.5: Grupos de Estados para las frecuencias y severidades de las cuatro coberturas.

En la Figura 3.5 se muestran los estados asignados mediante el método de clusterización a los grupos que se utilizarán como niveles de las variables categóricas en la aplicación de los modelos lineales generalizados, siendo la variable categórica **Grupo de Estado** con sus respectivos niveles (E1,E2,...,Em) para cada cobertura.

Las ligas que se utilizarón para las tres coberturas restantes para la frecuencia y la severidad son respectivamente:

1. Robo total: Liga Ward
2. Responsabilidad Civil: Liga Ward
3. Gastos Médicos: Liga Ward

Similarmente en la siguiente figura se muestra una tabla con las agrupaciones por carrocerías de las cuatro coberturas según frecuencia y severidad.

Agrupación de Carrocerías por Cobertura								
ESTADO	Daños Materiales		Robo Total		Responsabilidad Civil		Gastos Médicos	
	Frecuencia	Severidad	Frecuencia	Severidad	Frecuencia	Severidad	Frecuencia	Severidad
4 RUNNER	C1	C1	C1	C1	C1	C1	C1	C1
AVANZA	C1	C1	C1	C2	C2	C2	C2	C1
AVENSIS	C2	C2	C1	C3	C3	C3	C1	C1
CAMRY	C3	C1	C1	C2	C1	C1	C2	C1
COROLLA	C1	C1	C1	C2	C1	C2	C2	C1
FJ CRUISER	C3	C1	C2	C1	C4	C1	C1	C2
HIACE	C1	C1	C2	C1	C1	C2	C1	C2
HIGHLANDER	C3	C1	C1	C1	C1	C1	C1	C2
LAND CRUISER	C4	C1	C1	C2	C4	C2	C2	C1
MATRIX	C1	C2	C1	C2	C1	C3	C2	C1
MR2	C4	C1	C1	C3	C3	C4	C1	C2
PRIUS	C3	C1	C1	C3	C2	C3	C1	C2
RAV4	C1	C1	C1	C1	C1	C2	C2	C1
RUSH	C1	C1	C1	C3	C1	C3	C2	C1
SEQUOIA	C4	C3	C2	C1	C1	C1	C1	C2
SIENNA	C3	C1	C1	C1	C1	C2	C1	C1
SOLARA	C2	C4	C1	C3	C3	C5	C1	C1
YARIS	C1	C1	C1	C2	C1	C2	C2	C1

Figura 3.6: Grupos de Carrocerías para las frecuencias y severidades de las cuatro coberturas.

En la Figura 3.6 se muestran los carrocería asignados a los grupos que se utilizaran como niveles de las variables categóricas en la aplicación de los modelos lineales generalizados, siendo la variable categórica **Grupo de Carrocería** con sus repectivos niveles (C1,C2,...,Cn) para cada cobertura.

Las ligas que se utilizarón en las tres coberturas restantes para la frecuencia y la severidad son respectivamente:

1. Robo Total: Liga Ward
2. Responsabilidad Civil: Liga Ward y Liga Promedio
3. Gastos Médicos: Liga Ward

3.2. Estimación de los parámetros

Una vez ya teniendo definidas las variables categóricas y sus niveles tenemos por (2.14) que nuestro modelo para la frecuencia está dado por:

$$Fre(y_i) = e^{\beta_0+(\beta_{11}E1+\beta_{12}E2+...+\beta_{1m_1}Em_1)+(\beta_{21}C1+\beta_{22}C2+...+\beta_{2n_1}Cn_1)} \tag{3.1}$$

En este caso el intercepto β_0 es la variable offset más una constante.

Similarmente nuestro modelo para la severidad está dado de la siguiente manera:

$$Sev(y_i) = e^{\beta_0^*+(\beta_{11}^*E1+\beta_{12}^*E2+...+\beta_{1m_2}^*Em_2)+(\beta_{21}^*C1+\beta_{22}^*C2+...+\beta_{2n_2}^*Cn_2)} \tag{3.2}$$

Donde $y_i = (y_E, y_C)$ es un automóvil asegurado con y_E el estado de circulación y y_C su carrocería, por lo tanto de (3.1) y (3.2) se tiene que:

$$Ek = \begin{cases} 1 \quad \text{si} \quad y_E \in Ek \\ \\ 0 \quad \text{si} \quad y_E \notin Ek \end{cases} \qquad k = 1, 2, ..., m_j, \quad j = 1, 2$$

$$Cl = \begin{cases} 1 \quad \text{si} \quad y_C \in Cl \\ \\ 0 \quad \text{si} \quad y_C \notin Cl \end{cases} \qquad l = 1, 2, ..., n_j, \quad j = 1, 2$$

Como se mencionó anteriormente, la estimación de los parámetros $\beta's$ se hace por máxima verosimilitud, pero debido a la complejidad y que no hay métodos analíticos para estas estimaciones se emplea el paquete estadístico R para hacer dichas estimaciones a través de procedimientos numéricos.

Aplicando los modelos lineales generalizados a los grupos definidos anteriormente para la cobertura de daños materiaes, se tiene el siguiente resultado para la frecuencia:

```
Call:
glm(formula = FRE$NS ~ FRE$GE + FRE$GC, family = poisson(link = log),
    offset = log(FRE$EXPUESTOS))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0733  -0.6365  -0.0720   0.5799   2.3023

Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -1.87402    0.01355 -138.281  < 2e-16 ***
FRE$GE2     -0.13319    0.02755   -4.834 1.34e-06 ***
FRE$GE3      0.09701    0.01602    6.057 1.39e-09 ***
FRE$GE4      0.25214    0.01576   15.996  < 2e-16 ***
FRE$GE5      0.52204    0.02792   18.698  < 2e-16 ***
FRE$GE6      0.39382    0.10990    3.583 0.000339 ***
FRE$GC2     -1.00490    0.27753   -3.621 0.000294 ***
FRE$GC3     -0.07403    0.01184   -6.254 4.01e-10 ***
FRE$GC4     -0.21681    0.05698   -3.805 0.000142 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 754.132  on 22  degrees of freedom
Residual deviance:  29.003  on 14  degrees of freedom
AIC: 190.81

Number of Fisher Scoring iterations: 5
```

Figura 3.7: Resultados arrojados por R sobre el GLM para la frecuencia de la cobertura de Daños Materiales

Para el cálculo de estos parámetros se utilizó la distribución Poisson con la variable offset como el logaritmo de las unidades expuestas(Expuestos).

Como se observa en la Figura 3.7 los parámetros $\beta's$ todos son significativos, es decir, de la **Subsección 2.5.2** del capítulo anterior todos pasan el contraste de Wald, y por otro lado el valor del AIC lo podemos considerar como bajo respecto a los modelos con las distintas agrupaciones que se realizarón previamente, por tanto, podemos decir que este es un buen modelo para estimar la frecuencia.

De la misma manera se tiene el siguiente resultado para la aplicación de los modelos lineales generalizados para la severidad:

```
Call:
glm(formula = (SEV$SINOCU/SEV$NS) ~ SEV$GE + SEV$GC, family = Gamma(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.46830  -0.12651  -0.00207   0.07831   0.35504

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4747     0.1587  59.711 3.59e-15 ***
SEV$GE2       0.8358     0.1977   4.228  0.00142 **
SEV$GE3       0.4775     0.1977   2.416  0.03426 *
SEV$GE4       0.5440     0.1768   3.077  0.01052 *
SEV$GE5       0.8117     0.1977   4.106  0.00174 **
SEV$GE6       0.5893     0.1977   2.982  0.01249 *
SEV$GC2      -0.3423     0.1444  -2.371  0.03707 *
SEV$GC3       0.4786     0.1444   3.316  0.00688 **
SEV$GC4       0.8834     0.2205   4.006  0.00206 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.06251247)

    Null deviance: 5.33336  on 19  degrees of freedom
Residual deviance: 0.71293  on 11  degrees of freedom
  (3 observations deleted due to missingness)
AIC: 414.28

Number of Fisher Scoring iterations: 7
```

Figura 3.8: Resultados arrojados por R sobre el GLM para la severidad de la cobertura de Daños Materiales

Para el cálculo de estos parámetros se utilizó la distribución Gamma, al igual que en el caso de la frecuencia, para la severidad se observa en la Figura 3.8 que los parámetros $\beta^{*\prime}s$ todos son significativos y aunque en este caso el valor del AIC es un poco más alto que para el caso de la frecuencia la devianza es más baja en este modelo.

A continuación se presentan los parametros estimados de la frecuencia y de la severidad de las cuatro coberturas.

Parámetro	Daños Materiales		Robo Total		Responsabilidad Civil		Gastos Médicos	
	Frecuencia	Severidad	Frecuencia	Severidad	Frecuencia	Severidad	Frecuencia	Severidad
β_0	-1.87402	9.4747	-7.4838	11.4518	-2.99559	9.61168	-4.69048	8.8175
β_{11}	0	0	0	0	0	0	0	0
β_{12}	-0.13319	0.8358	2.0866	0	-0.51088	0.34164	-0.67333	-0.6102
β_{13}	0.09701	0.4775	1.2118	0.52502	0.13784		0.36422	0.864
β_{14}	0.25214	0.544	2.5037	1.18521	-0.18846		1.14349	
β_{15}	0.52204	0.8117	3.075	0.78896	0.43327			
β_{16}	0.39382	0.5893	3.3416					
β_{21}	0	0	0	0	0	0	0	0
β_{22}	-1.0049	-0.3423	0.8761	-0.33691	-0.23554	-0.10664	0.67875	0.7732
β_{23}	-0.07403	0.4786		0	-0.61737	-0.35053		
β_{24}	-0.21681	0.8834			0.31178	1.24923		
β_{25}						0.27698		

Figura 3.9: Parámetros estimados para la frecuencia y severidad modeladas de las cuatro coberturas.

Como se observa en la Figura 3.9, los parámetros asignados a los primeros grupos de estados y carrocerías son todos cero, esto se debe a que cuando la variable explicativa es categórica como en este caso, el modelo se construye considerando variables numéricas asociadas a la categórica, son las llamadas variables de diseño o auxiliares.

Cuando se tienen variables categóricas con más de dos niveles, digamos con n niveles, se construyen $n-1$ variables de diseño; R toma el primer nivel de cada categoría como referencia, de manera que todas las variables de diseño toman el valor 0 para esa categoría, es decir los grupos E1 y C1 para la frecuencia y severidad toman el valor 0 por tanto los valores de β_{11}, β_{21}, β^*_{12} y β^*_{21} son cero.

3.3. Cálculo de la prima de Riesgo

Una vez estimados los parámetros de la frecuencia y la severidad, el cálculo de la prima de riesgo, a la que llamaremos prima modelada o prima estimada, queda como lo establece la ecuación (2.16). Los resultados de las primas modeladas se pueden observar en los **Anexos.**

En esta sección, se presentan los comparativos de las primas de riesgo modeladas y las primas de riesgo que se calculan multiplicando la frecuencia (número de siniestros entre unidades expuestas) por la severidad (monto de siniestro entre unidades expuestas), a la que llamaremos primas estadísticas. Con fines de apreciación, las gráficas comparativas se realizan en escala logarítmica donde el eje x es un ordena-

miento alfabético por estado y carrocería de las primas puras de riesgo.

Antes de presentar los comparativos, se explicará brevemente la forma en que se valida que las primas modeladas arrojan una estimación adecuada, que no subestima o sobrestima el siniestro ocurrido total de la cartera.

3.3.1. Comparativo

Como se mencionó con anterioridad, el cálculo de las primas de riesgo a través de regresiones lineales generalizadas, genera estimaciones con menor volatilidad, que el método comúnmente usado de multiplicar la frecuencia (número de siniestros entre unidades expuestas) por la severidad (monto de siniestro entre unidades expuestas); de esta forma, un comparativo directo entre ambos métodos, si bien resulta ilustrativo, no es adecuado para determinar si la prima modelada estimada adecuadamente el riesgo de la cartera, pues como se verá posteriormente, puede existir mucha variabilidad entre algunas primas modeladas y algunas primas estadísticas.

Por tal motivo, para determinar que no hay una subestimación o sobrestimación de las primas de riesgo, se propone comparar el siniestro ocurrido total $Sinocu_T$ o costo total c_T de las reclamaciones de la base[4](segmentadas por cobertura y considerando únicamente carrocerías TOYOTA) con el siniestro ocurrido total estimado $\widehat{Sinocu_T}$ o costo total estimado $\widehat{c_T}$ que se genera al sumar el producto de la prima modelada $\widehat{PR}$ con sus respectivos expuestos:

$$\widehat{Sinocu_T} = \widehat{c_T} = \sum \widehat{PR} * expuestos$$

De esta forma, al comparar c_T de la cartera, contra $\widehat{c_T}$ estimado por las primas modeladas, se puede determinar una tolerancia o nivel de significancia que proporcione un indicador de la precisión del modelo. Este indicador α, será el porcentaje de error del costo total estimado $\widehat{c_T}$ y el costo total c_T de la cartera (segmentada por cobertura), quedando como:

$$\alpha = |\frac{\widehat{c_T}}{c_T} - 1| * 100\,\%$$

Por hábito, estableceremos que la prima de riesgo modelada o estimada es adecuada si el nivel de tolerancia es menor a $5\,\%$.

[4]Al provenir de la base, también lo denotaremos como siniestro ocurrido estadístico o costo estadístico.

3.3.2. Daños Materiales

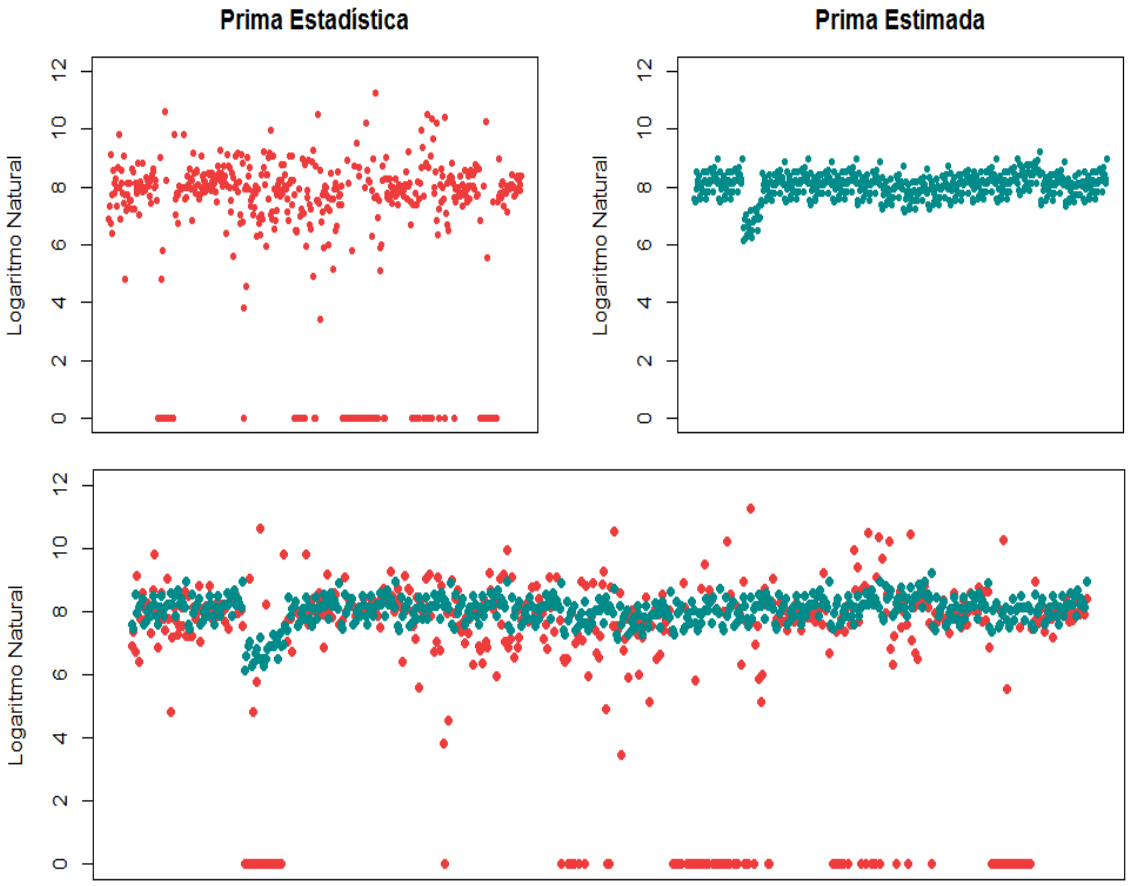


Figura 3.10: Prima Riesgo modelada de la cobertura de Daños Materiales, donde el eje x es un ordenamiento alfabético por estado y carrocería.

En la gráfica de la prima estadística (superior izquierda) se observa una variabilidad muy grande además de que se observan puntos con valor cero que son las carrocerías que no reportaron siniestros o que el costo del siniestro fue nulo, por lo que se tiene el problema de cuanto se le tiene que cobrar a las carrocerías con características similares, es por eso que este método no es adecuado para establecer las cuotas a cobrar.

En la gráfica de la prima estimada (superior derecha) se observa un comportamiento más homogéneo, es decir, los datos que en la gráfica de la prima estadísticas hacían que hubiese una variabilidad grande ahora se ajustaron con el comportamiento de carrocerías con características similares, es por eso que con este modelo se puede definir la cantidad que se va a cobrar por el cargo de la cobertura según la carrocería y estado de circulación.

En la tercera gráfica se puede observar mejor la comparación de ambos métodos, y cómo es que los datos que hacen que la prima estadística tenga una variabilidad

grande en la prima estimada se ajustan mejor al modelo y disminuyen esa variabilidad.

A continuación se mostrará gráficamente el costo estadístico y el costo estimado, esto para tener una perspectiva de las garantías que tendrán tanto el asegurado como la aseguradora que el cobro de la prima de riesgo para ambas partes es lo más justo.

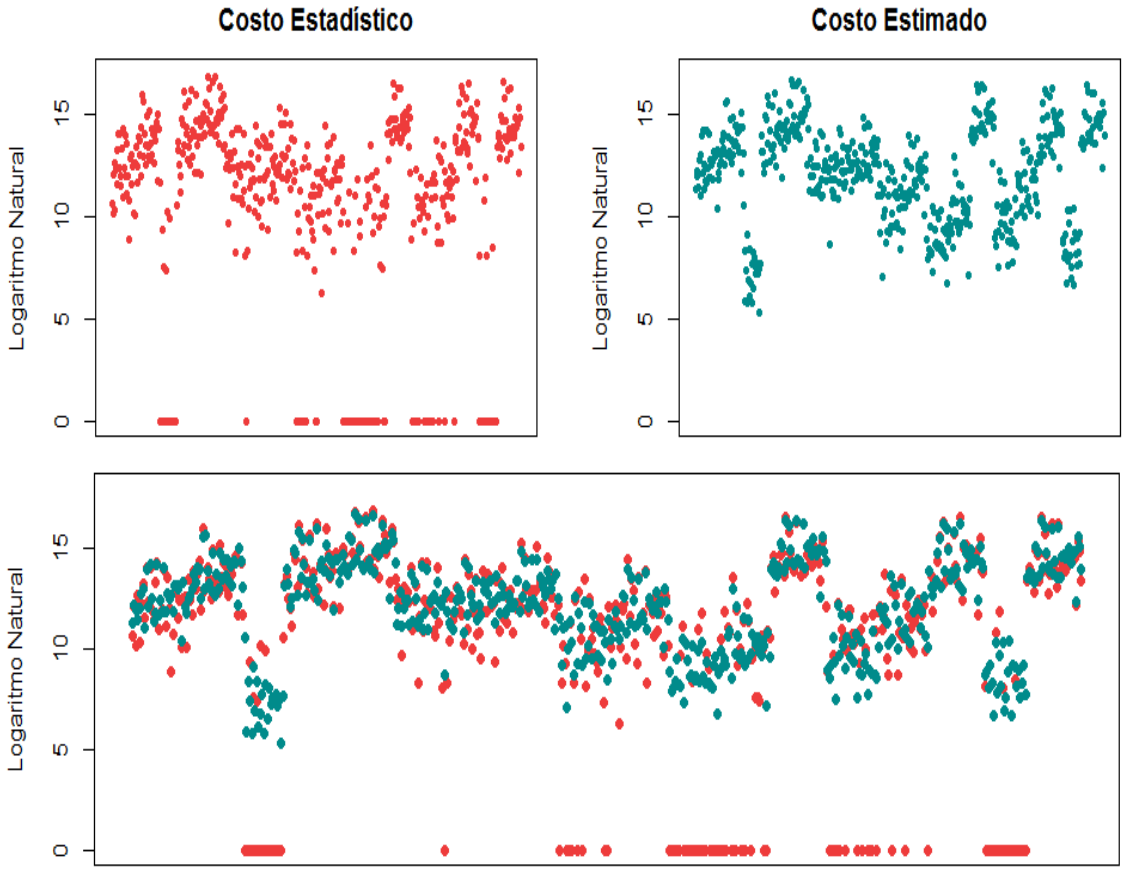


Figura 3.11: Siniestro Ocurrido o Costo Total de la cobertura de Daños Materiales, dondel eje x es un ordenamiento alfabético por estado y carrocería.

En la figura del costo estadístico (superior izquierda) se vuelven a observar las carrocerías que no tuvieron siniestros o el costo fue nulo, por tanto al hacer la estimación de dicho costo (superior derecha) dichas estimaciones se ajustan a las demás, además se puede observar que la estimación tiene un comportamiento muy similar al estadístico lo que nos indica que nuestro modelo puede ser confiable. En la figura de abajo, se muestra la comparación de ambos costos y quitando las observación cero se puede ver que no hay mucha variabilidad entre ambas.

Con este modelo se obtiene un indicador α de 2.91 % lo que nos indica que es menor al 5 %, por tanto podemos considerar una buena estiamación para la prima pura de riesgo.

3.3.3. Robo total

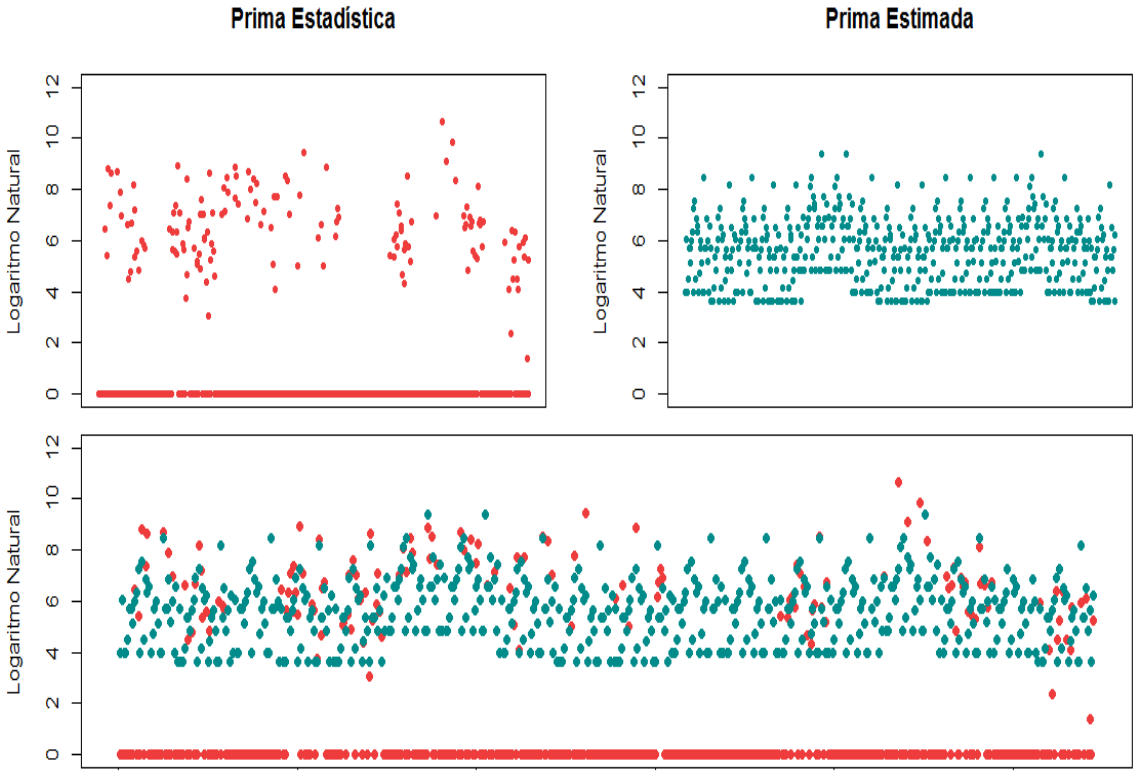


Figura 3.12: Prima Riesgo modelada de la cobertura de Robo Total, donde el eje x es un ordenamiento alfabético por estado y carrocería.

En la gráfica de la prima estadística (superior izquierda) al igual que la cobertura anterior se observa una variabilidad muy grande y también se observan más puntos con valor cero, esto nos indica que en esta cobertura el número de siniestros es muy bajo por lo mismo es un poco más difícil tener un criterio apropiado para el cálculo de la prima de riesgo.

En la gráfica de la prima estimada (superior derecha) se observa como es que dependiendo de la carrocería y el estado de circulación a los autos que no tienen siniestro reportado se les asigna la prima de los que si tuvieron al menos un sinietro reportado.

En la tercera gráfica observamos mejor la comparción de ambas, viendo como la mayoría de las primas estimadas son menores a las estadísticas.

Ahora veremos si la estimación de dicha prima que se propone garantiza a la aseguradora poder cubrir los costos de los siniestros que se esperan.

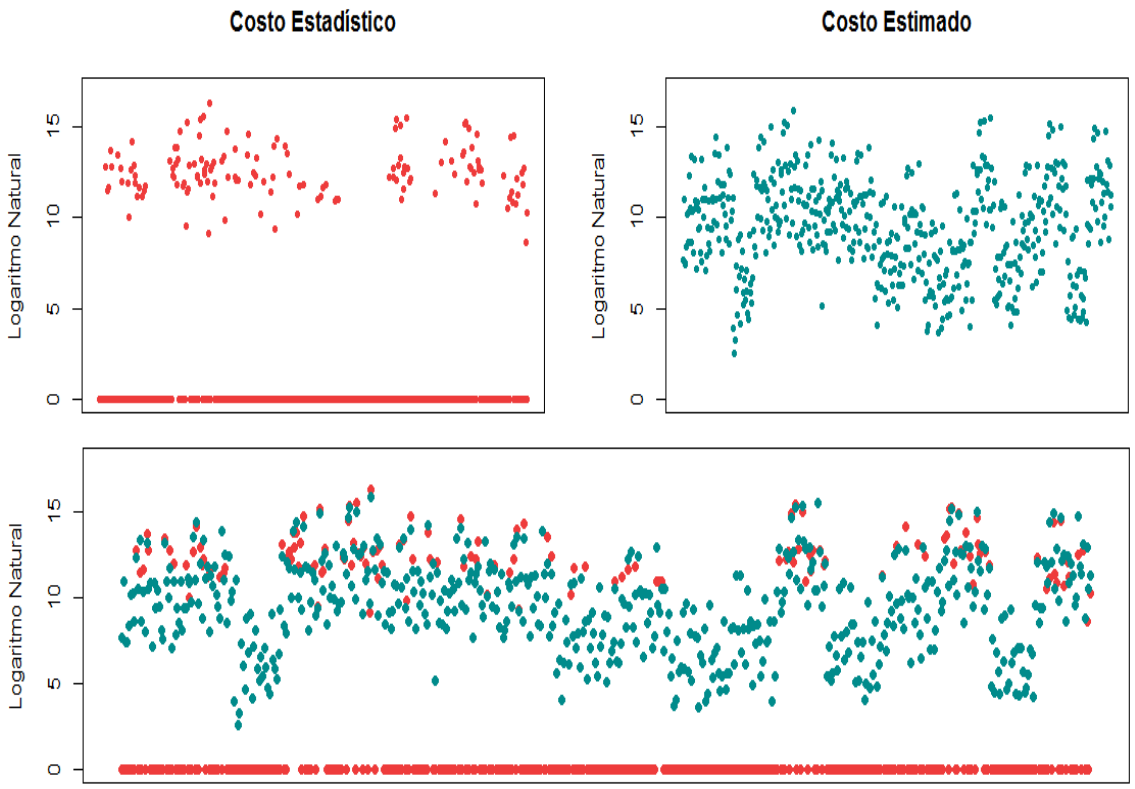


Figura 3.13: Siniestro Ocurrido o Costo Total de la cobertura de Robo Total, donde el eje x es un ordenamiento alfabético por estado y carrocería.

En estas gráficas se observa que la distribución de los costos estadísticos es alta ya que se tiene que cubrir el precio del auto robado, es por eso que los costos estimados tienen una distribución diferente ya que contemplan a todos los autos asegurados.

Con este modelo se tiene una significancia de 1.31 % lo que nos indica que es menor que $\alpha = 5\,\%$, por tanto podemos considerar una buena estimación para la prima pura de riesgo de la cobertura de Robo Total, sin presentarse una subestimación ni sobrestimación el costo total.

3.3.4. Responsabilidad Civil

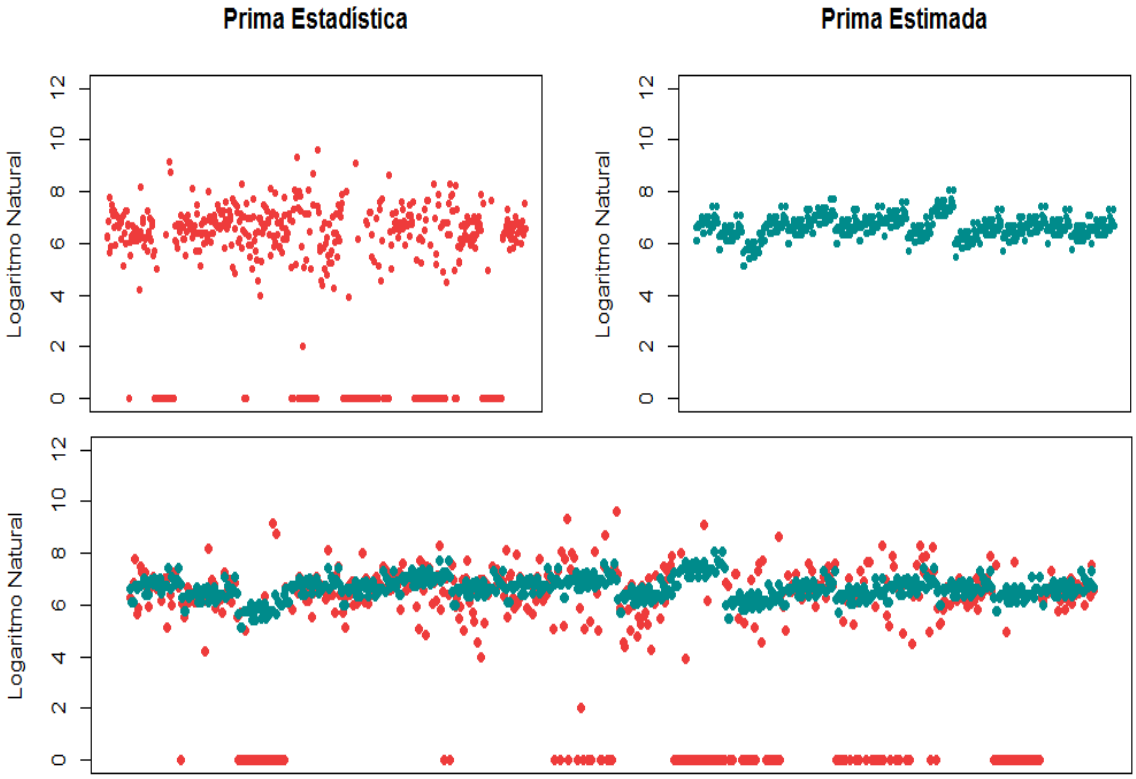


Figura 3.14: Prima Riesgo modelada de la cobertura de Responsabilidad Civil, donde el eje x es un ordenamiento alfabético por estado y carrocería.

Para esta cobertura se obseva que la variabilidad de la prima estadística no es tan grande como las anteriores, por eso la prima estimada tienen una distrubución más uniforme.

A pesar de que se tiene poca variabilidad, es necesario ver si existe una subestimación o sobrestimación del siniestro ocurrido total.

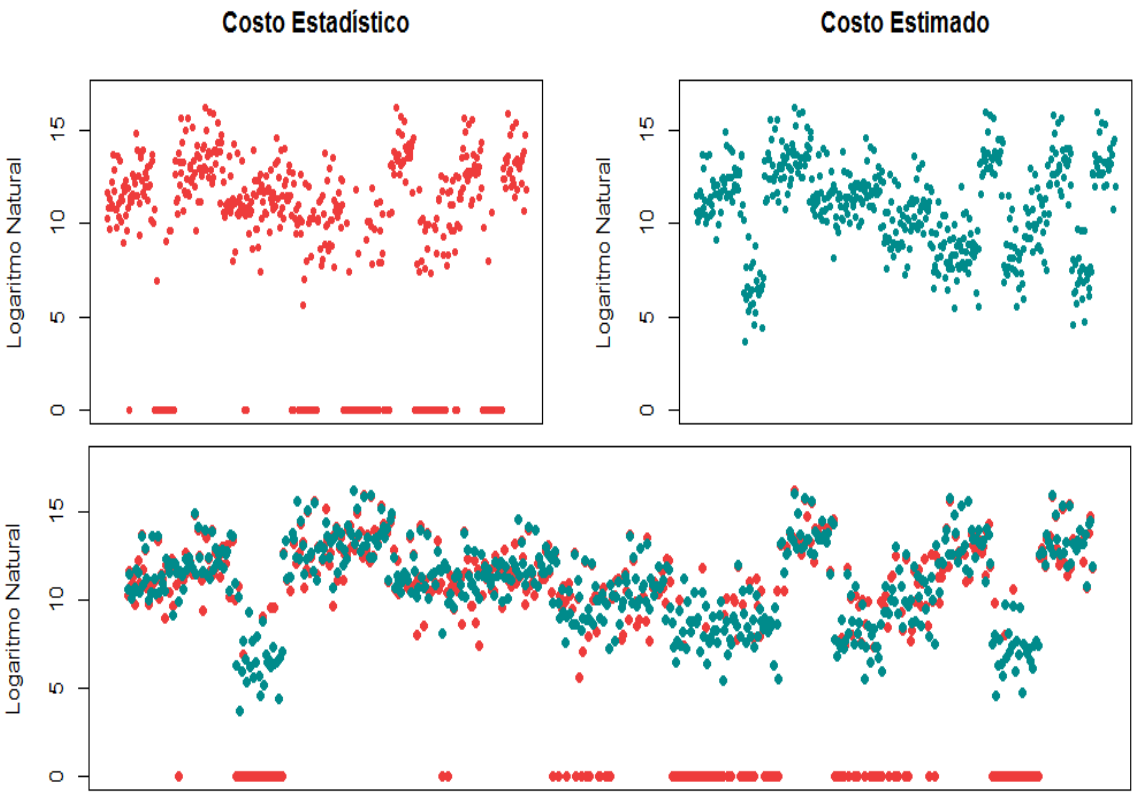


Figura 3.15: Siniestro Ocurrido o Costo Total de la cobertura de Responsabilidad Civil, donde el eje x es un ordenamiento alfabético por estado y carrocería.

En estas gráficas se observa cómo está la distribución de los costos; a diferencia de las primas si tienen una variabilidad grande, aun así los comportamientos de ambos costos son muy similares.

Con este modelo aunque haya una variabilidad grande en los costos se tiene una significancia de 0.99 % lo que nos indica que es menor que $\alpha = 5\,\%$, por tanto podemos considerar que se ha generado una buena estimación para la prima pura de riesgo de la cobertura de Responsabilidad Civil.

3.3.5. Gastos Médicos

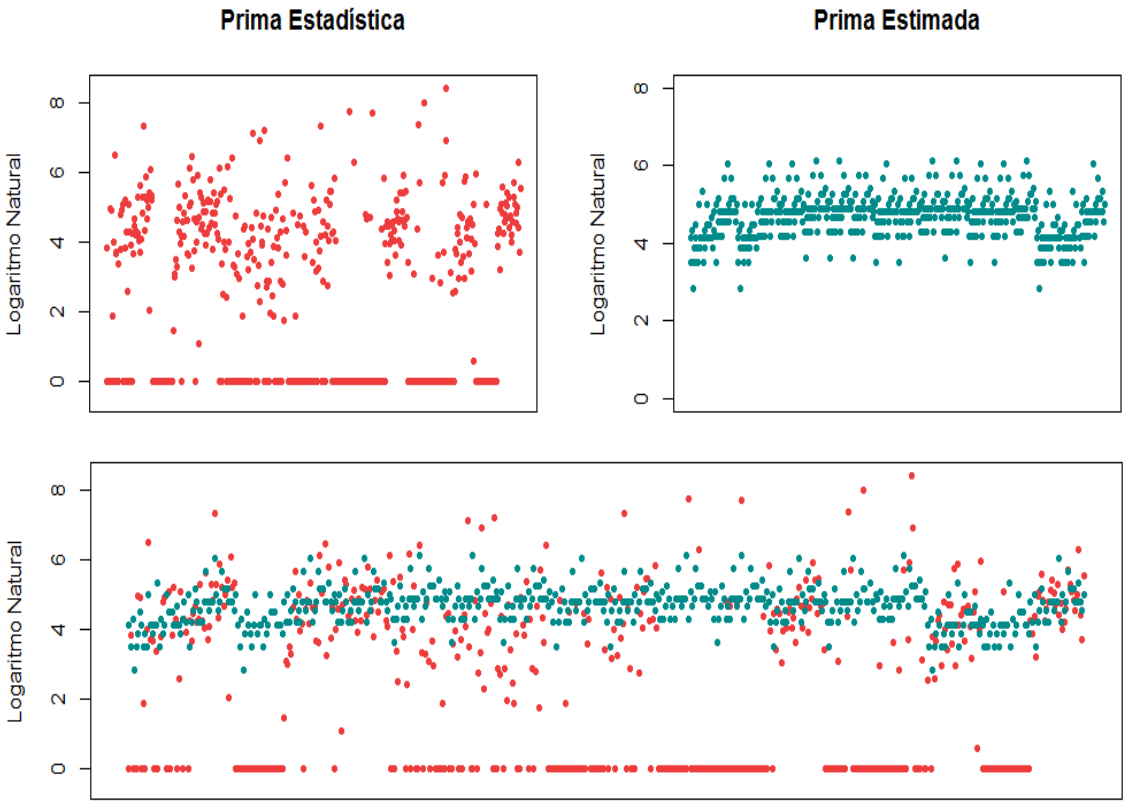


Figura 3.16: Prima Riesgo modelada de la cobertura de Gastos Médicos, donde el eje x es un ordenamiento alfabético por estado y carrocería.

En esta cobertura, a diferencias de las anteriores, las primas tiene una variabilidad más grande presentándose muchos puntos en cero de la prima estadística, en al gráfica de abajo se ve como las primas estadísticas están alrededor de la estimada.

Ahora veremos el comportamiento de los costos de esta cobertura para ver si existe una subestimación o sobrestimación por parte del modelo.

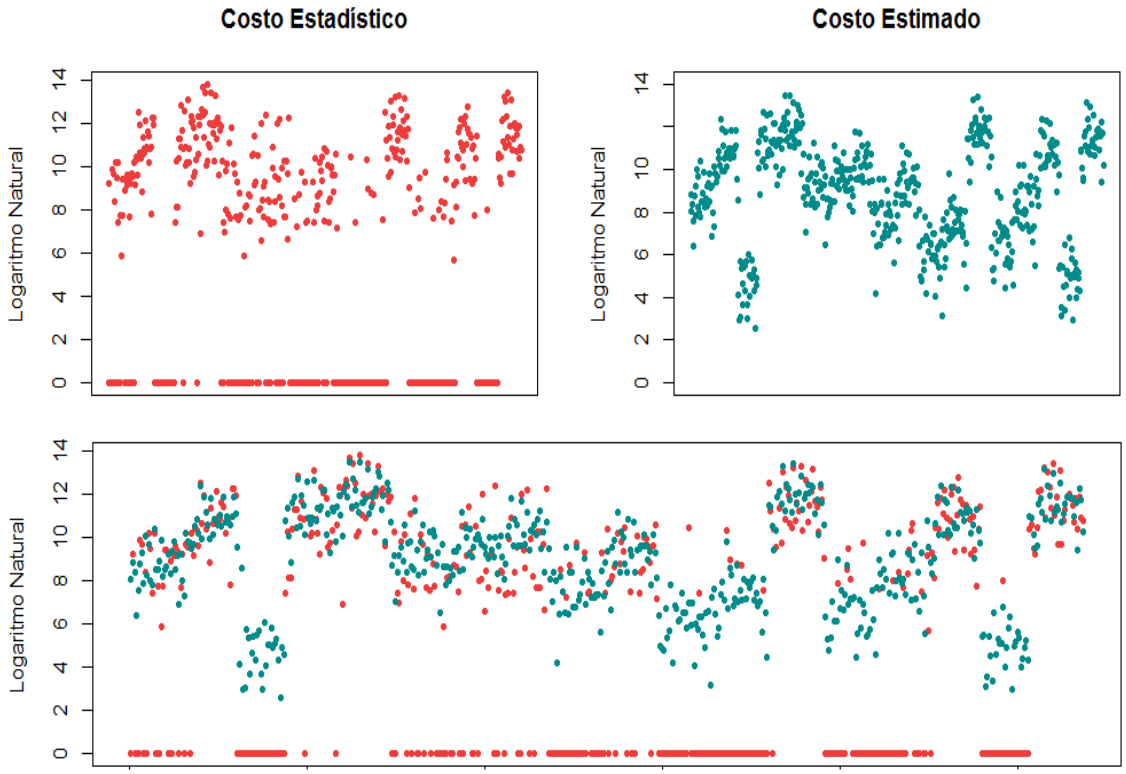


Figura 3.17: Siniestro Ocurrido o Costo Total de la cobertura de Gastos Médicos, donde el eje x es un ordenamiento alfabético por estado y carrocería.

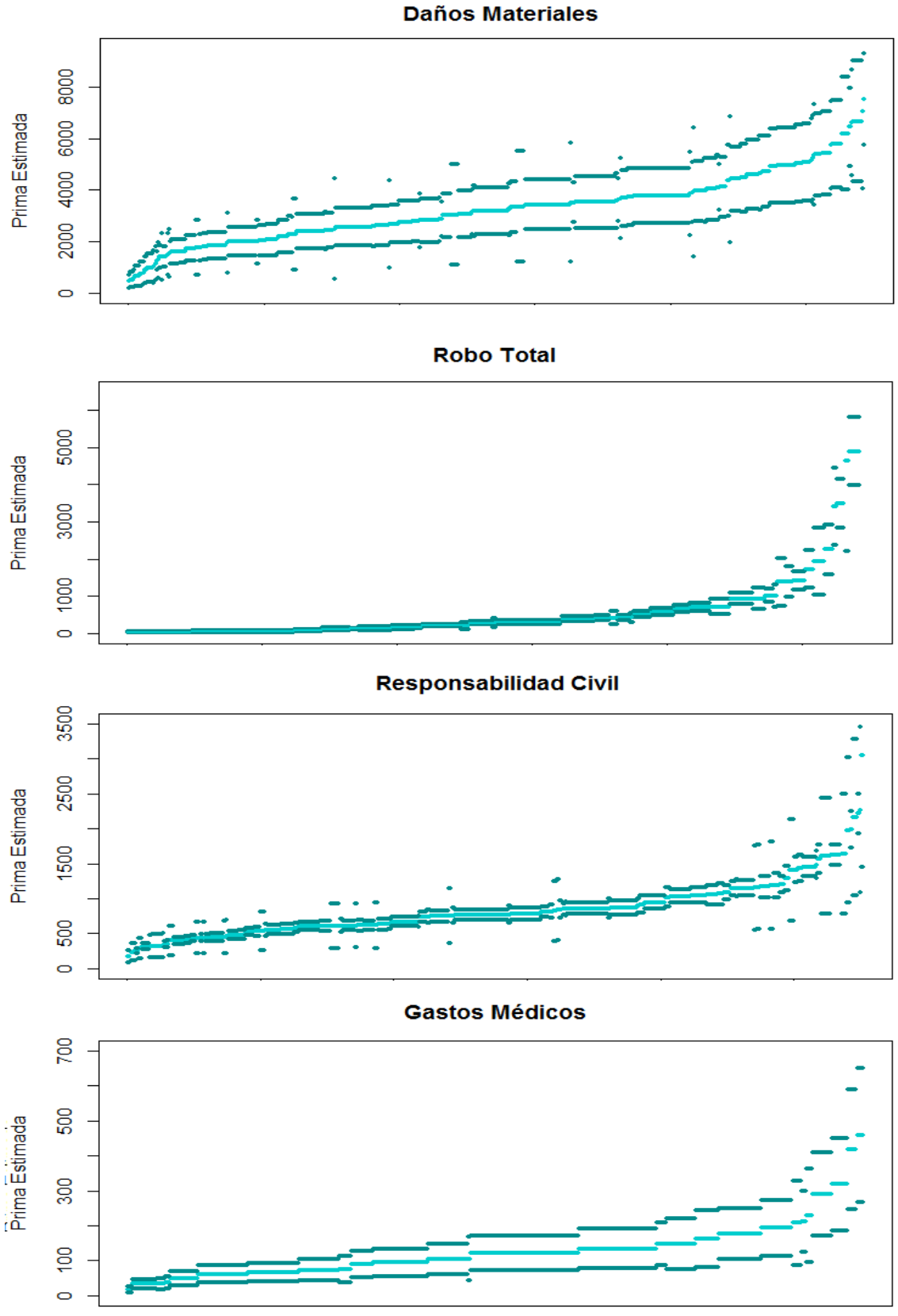
Con este modelo se tiene una significancia de 0.95 % que a pesar de que es la más grande de las cuatro coberturas, nos indica que es menor que $\alpha = 5\,\%$, por tanto podemos considerar que el modelo genera una buena estimación del siniestro ocurrido total.

3.3.6. Intervalos de Confianza de las primas de riesgo

Una vez estimadas las primas de riesgo para las cuatro coberturas y de haber verificado que el costo del siniestro estadístico y estimado estén dentro de la tolerancia establecida, el siguiente paso es tener un cierto margen de confiabilidad de la estimación de las primas de riesgo, esto lo haremos calculando sus intervalos de confianza para cada cobertura. De la **sección 2.6** del capítulo anterior tenemos que el intervalo propuesto para las primas de riesgo está dada por (2.20), por lo que debemos encontrar los errores estándar de las frecuencias y severidades estimadas, esto lo haremos evaluando las frecuencias y severidades estimadas en los modelos lineales generalizados obtenidos para cada cobertura, para así poder obtener sus errores estándar y así poder aplicar (2.20).

A continuación se muestran las primas de riesgo a su escala normal con su respectivo intervalo de confianza a un 95 %, para poder tener una mejor visualización se ordenaron las primas de riesgo de menor a mayor.

Intervalos de Confianza



En las gráficas anteriores se puede observar que en la cobertura de Daños Materiales hay intervalos con una variabilidad grande, esto se puede deber a que hay grupos con pocos elementos y al tener poca información sus errores estándar son más grandes que otras primas donde sus estados y carrocerías pertenecen a grupos con más elementos.

Como los datos son de un solo año, esperaríamos que al siguiente año los siniestros reportados tengan un comportamiento similar, es por eso la importancia de los intervalos de confianza ya que nos da la probabilidad que contengan al valor estimado de la prima de riesgo del siguiente año.

3.3.7. Resultados Generales

Para tener una idea más clara de cuáles son los Estados y las Carrocerías que presentan un mayor riesgo, se muestra a continuación un comparativo gráfico del promedio de la prima de riesgo por cobertura, cabe señalar que la escala no es la misma entre las coberturas ya que el objetivo de la gráficas es ver en que estados y carrocerías la prima de riesgo es más baja y más alta.

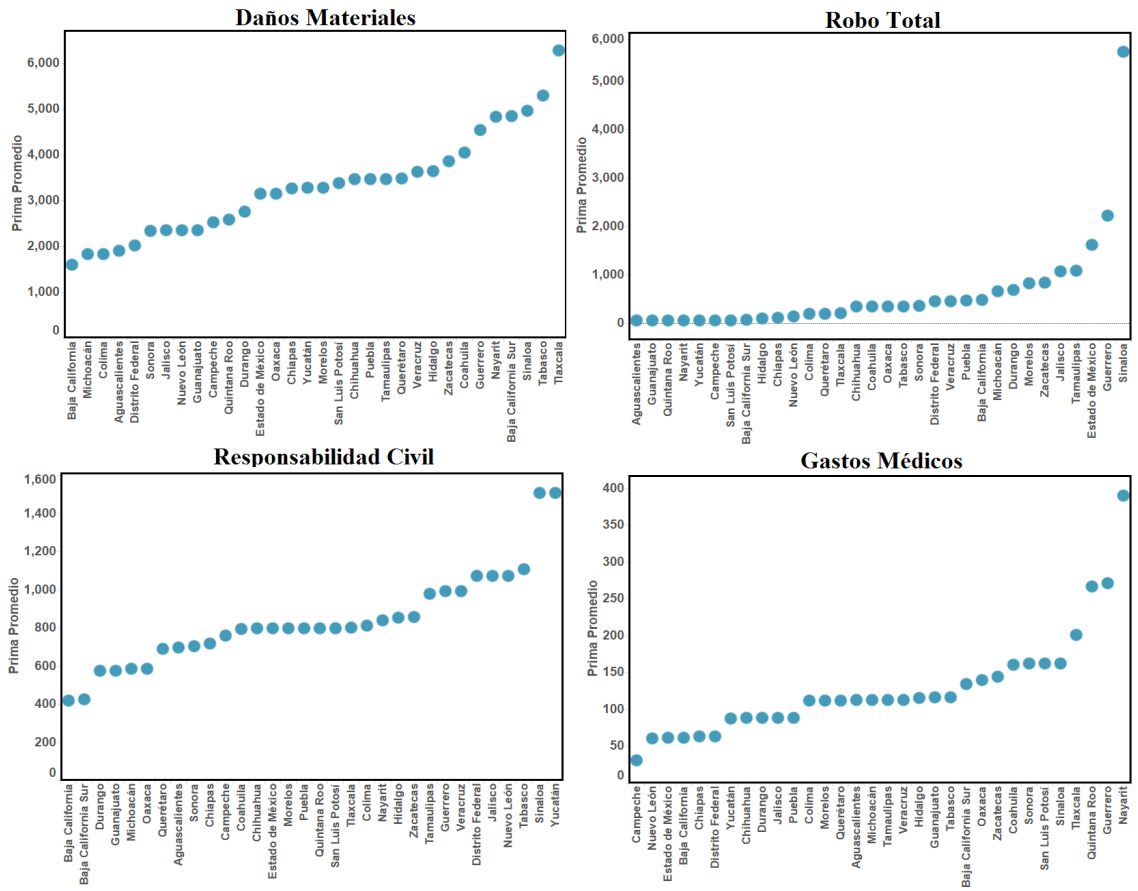


Figura 3.18: Primas de Riesgo Promedio en los Estados para las cuatro coberturas.

En las gráficas anteriores se pueden observar que el orden de los estados va va-

riando según el tipo de cobertura que se esté analizando, es decir, algunos estados presentan una prima de riesgo más alta para algunas coberturas y en otras es más baja en comparación con otros estados.

Similarmente ahora se presenta el comparativo de la prima de riesgo por tipo de carrocería.

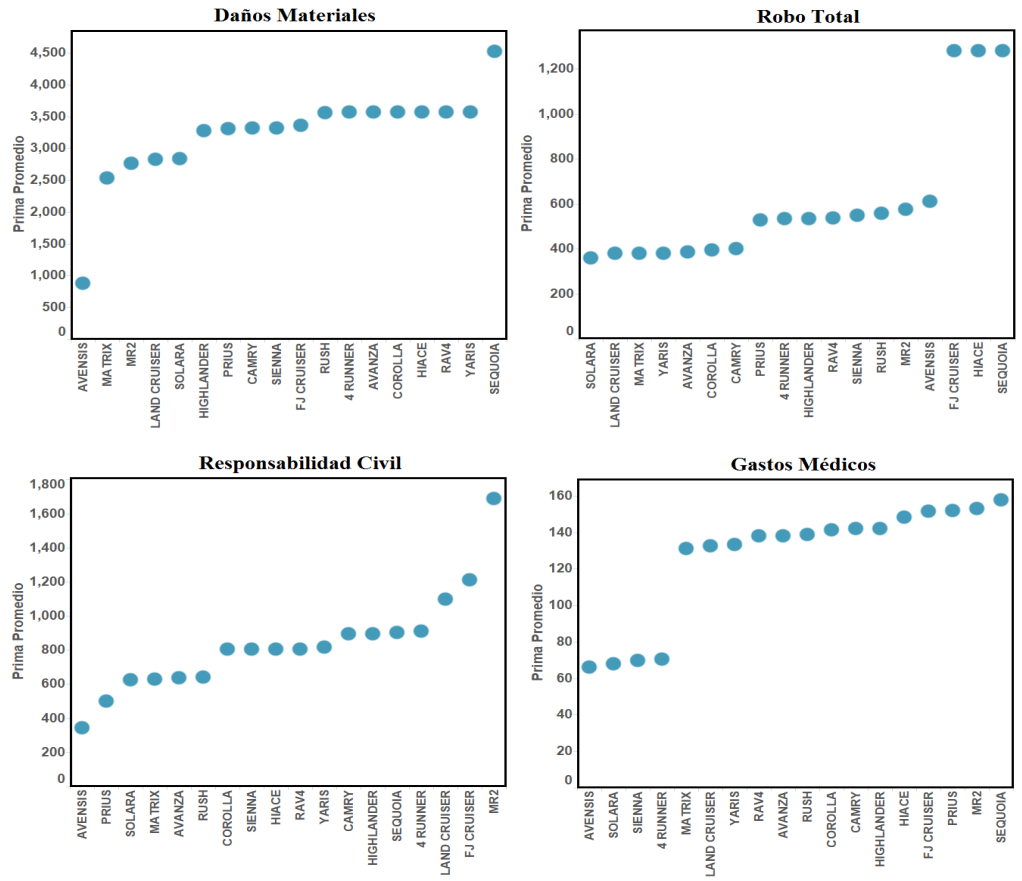


Figura 3.19: Primas de Riesgo Promedio por tipo de Carrocería para las cuatro coberturas.

Al igual que para los estados, las gráficas anteriores muestran las carrocerías con un mayor riesgo según la cobertura; se observa que el ordenamiento de las carrocerías varía según la cobertura.

Conclusiones

Con base en el objetivo del presente trabajo, se llegó a las siguientes conclusiones.

La prima pura de riesgo es la prima que se obtiene en función de la probabilidad de que ocurra un siniestro y el monto de la pérdida que tendría el asegurado, la cual tiene como objetivo cubrir los costos de las posibles reclamaciones que se puedan presentar a lo largo de un contrato; para el caso del presente trabajo, se obtuvo que las estimaciones de dichas primas de las cuatro coberturas (Daños Materiales, Robo Total, Responsabilidad Civil y Gastos Médicos) estuvieron dentro de una tolerancia planteada menor a un 5 % del siniestro ocurrido o costo total estadístico, evitando una subestimación o sobrestimación, permitiendo un margen adecuado para establecer que las primas serán suficientes para afrontar las obligaciones contractuales de pagos de siniestros.

Los resultados presentados se basaron en la metodología propuesta en este trabajo, en la cual se toman en cuenta algunos factores de riesgo asociados a los asegurados que influyen en el comportamiento siniestral y que intervienen en el proceso de tarificación. Debido a la diversificación de estos factores fue necesario hacer una segmentación teniendo en cuenta su experiencia siniestral en grupos homogéneos de manera que los asegurados que pertenecen a un mismo grupo paguen la misma tarifa ya que con esto se evitó la antiselección; para hacer esta segmentación se utilizaron algunos de los métodos aglomerativos de clusterización vistos en la **Sección 1.3** del **Capítulo 1**. En la aplicación de dichos métodos se utilizaron varias ligas y se tomaron en cuenta diferentes números de grupos para poder tener una mayor comparación y así poder elegir la segmentación más idónea para el cálculo de la prima de riesgo. Para el caso de la frecuencia no hubo mayor problema ya que se modeló con una distribución Poisson, pero para el caso de la severidad, que se modeló con una distribución Gamma, que es para valores estrictamente positivos, los estados o carrocerías con siniestro ocurrido cero se agruparon con los estados y carrocerías con siniestros ocurridos más bajos formando grupos con más estados y carrocerías y disminuyendo el número de grupos.

La parte crucial de este trabajo fue la aplicación de los modelos lineales generalizados, ya que las variables explicativas utilizadas no fueron numéricas sino

categóricas donde estas categorías fueron los resultados obtenidos en el proceso de agrupación.

Como se mencionó anteriormente, se realizaron varios números de grupos para los Estados y las Carrocerías, así se pudo aplicar los modelos lineales generalizados con todas las combinaciones posibles entre los grupos de Estados y Carrocerías para tener una mayor comparación y seleccionar el que mejor cumpliera con los criterios de selección del modelo vistos en la **Sección 2.5** del **Capítulo 2** como el test de Wald para los estimadores, la devianza y el criterio de información AIC; durante este proceso se observó que los modelos para la frecuencia en su mayoría no presentaron mayor dificultad y sólo se seleccionaron los que tuvieran menor devianza y menor AIC comparado con otros modelos con distintos número de grupos, pero para el caso de la severidad se presentaron más dificultades ya que en algunos casos los estimadores no pasaban el test de Wald, que es el primer criterio para la selección del modelo; en el caso de la cobertura de Gastos Médicos solo se obtuvo un modelo para la severidad el cual paso el test de Wald para los estimadores, a diferencia de las otras tres coberturas donde se obtuvo más de un modelo que cumplieron con este criterio.

Para el cálculo de la prima de riesgo se tuvo que hacer el producto de la frecuencia por la severidad de todas las combinaciones de los modelos que se tenían, por lo que los modelos presentados fueron los que obtuvieron una menor tolerancia de acuerdo a lo planteado en la **Sección 3.3.1** del **Capítulo 3**, de ahí la importancia de haber realizado varias agrupaciones para obtener varios modelos y elegir la que mejor se adecuara al objetivo planteado.

Un factor importante en la realización de este trabajo fue la visualización de los modelos en estudio, ya que permitió dar una perspectiva no solo numérica sino también gráfica de la modelación de la prima de riesgo en seguros de automóviles, ya que en las gráficas presentadas se observa la comparación entre la forma común de calcular la prima de riesgo y la metodología basada en la combinación de clusterizacion y los modelos lineales generalizados, también se da una perspectiva del comportamiento de la prima por estado o por carrocería; por otra parte muestra los puntos en los que se debe seguir trabajando para mejorar, como es el caso de la cobertura de Daños Materiales, que además de ser la cobertura que tiene una tolerancia mayor, sus intervalos de confianza presentan una mayor longitud para cada estimación, esto se puede deber a la poca información que se tienen en las agrupaciones propuestas debido a la gran diversificación en que esta cobertura pueda ser requerida.

Los resultados obtenidos cumplen con el objetivo de este trabajo, ya que se presentó toda una metodología para el cálculo de la prima de riesgo en seguro de automóviles y arrojan para algunas coberturas buenas estimaciones y para otras como el caso de Daños Materiales, que no eran buenas pero si aceptables, ya que se evitó la subestimación y sobrestimación. Una propuesta para mejorar los modelos propuestos en este trabajo, podría ser tomar en cuenta más factores relevantes para

ser consideradas como categorías en la aplicación de los modelos lineales generalizados, por ejemplo, el modelo del Automóvil, la historia siniestral del asegurado, entre otras.

Con el desarrollo de este trabajo se obtienen algunas contribuciones considerables; dado que el tema de la prima de riesgo en seguro de automóviles es muy importante para el sector asegurador, se estableció una metodología completa con una combinación de métodos estadísticos multivariados que pueden ser utilizados como un parámetro de comparación a los resultados que presentan cada una de las compañías aseguradoras ya que el buen cálculo de ésta prima les permitirá cumplir con las demandas pactadas a las que sean sujetas sin tener que elevar sus reservas, por otra parte, este cálculo impactaría directamente en la tarifa comercial ofertando tarifas competitivas dentro de todo sector asegurador; una segunda contribución fue el presentar un intervalo de confianza para las estimaciones de las primas, este intervalo permite un margen de error con el cual se podría establecer las reservas necesarias y obtener una solvencia suficiente para cumplir con mayor eficiencia y sin problemas las posibles reclamaciones que se presenten.

Durante el desarrollo de este trabajo se presentaron algunas limitaciones, ya que por falta de tiempo no se pudo analizar todas las carrocerías de todas las marcas de automóviles, analizando solo las carrocerías de la marca TOYOTA, así, si alguna compañía adoptara esta metodología tendría que hacer análisis para cada una de las marcas de automóviles de manera independiente lo que implicaría un mayor trabajo en la obtención de tarifas suficientes para todo su catálogo de marcas aseguradas; en cuestión del manejo de los datos se llevó más tiempo de lo que se tenía planeado, ya que al tener observaciones con cero números de siniestros hacían más complicado encontrar los grupos ideales que cumplieran con los supuestos estadísticos en la aplicación de los Modelos Lineales Generalizados, ya que para la severidad se tuvieron que hacer menos grupos con más miembros (Estados y Carrocerías) para poder obtener una severidad grupal estrictamente positiva debido a que se modelo con una distribución Gamma.

Finalmente, se deja un punto de partida para seguir trabajando y mejorar el cálculo de la prima pura de riesgo, una posible extensión de este trabajo es la modelación mediante copulas bivariadas [10], en vez de modelar la frecuencia y la severidad de manera independiente como se hizo, se puede construir un modelo conjunto que permita explícitamente una dependencia entre el costo del siniestro y el número de siniestros, combinando distribuciones marginales para la frecuencia y severidad y así obtener una función de distribución de probabilidad conjunta. El supuesto de independencia limita el cálculo de la prima de riesgo, es por eso que al encontrar un modelos donde la frecuencia y la severidad se correlacionen podrían minimizar aún más la variabilidad presentada; el modelo de copulas bivariadas asume que las distribuciones marginales pueden ser una combinación de funciones de probabilidad tanto discretas como continuas es por eso la importancia de esta posible alternativa a la metodología presentada en este trabajo.

Bibliografía

[1] Armstrong David A., ***Factorplot: Improving Presentation of Simple Contrast in Generalized Linear Models.*** The R Journal Vol 5/2 December.

[2] Avella Medina Marco and Ronchetti Elvezio, ***Robust and consisten variable selection for generalized linear and additive models.*** Research Center for Statistic and Genova School of Economics and Management, University of Genova, Switzerland, 2014.

[3] Christopher R. Bilder, Thomas M. Loughin., ***Analysis of Categorical Data with R.*** Chapman & Hall/CRC Texts in Statical Science, 2004.

[4] Dobson Annette J., ***An Introduction to Generalized Linear Models.*** Chapman & Hall, Second Edition, 2002.

[5] Everitt Brian, Hothorn Torsten., ***An Introduction to Applied Multivariate Analysis with R.*** Springer, 2011.

[6] Everitt Brian, Landau S,. Leese M., and Stahl. ***Cluster Analysis.*** Chichester, UK: John Wiley & Sons, 5thh Edition, 2011.

[7] Faraway Julian L., ***Linear Models with R.*** Chapman & Hall/CRC Texts in Statical Science, 2004.

[8] Faraway Julian L.,***Extending the linear models with R: Generalized Linear, Mixed effects and Nonparametric Regression Models.*** Taylor & Francis Group, 2006.

[9] Kass Rob, Goovaerts Marc, Dhaene Jan, Denuit Michel., ***Modern Actuarial Risk Theory Using R.*** Second Edition, Springer, 2008.

[10] Krämer Nicole, Brechmann Eike C., Silvestrini Daniel, Czado Claudia., ***Total loss estimation copula-based regression models.*** Journal Insurance: Mathematics and Economics, TU München, Department of Mathematical Statistics, Parking 13, 85748 Garching Germany, 2013.

[11] Peña Daniel, ***Análisis de Datos Multivariantes.*** Primera Edición, Mc Graw Hill, 2002.

[12] Piet de Jong and Gillian Z. Heller., ***Generalized Linear Models for Insurance Data.*** Cambridge Univesity Press, 2008.

[13] Shmueli Galit, Patel Nitin and Bruce Peter., ***Data mining for Business Intelligence, Concepts, Technics and Applications.*** Wiley, Second Edition, 2011.

[14] Stroup Walter W., ***Generalized Linear Mixed Models: Modern Concepts, Methods and Applications.*** CRR Press, Taylor and Francis Group, 2012.

[15] https://www.r-project.org/, ***The R Project for Statistical Computing.***

[16] http://www.segurb2b.com/informacion/dicc_seguros.cfm.

Anexos

En esta sección se presentán las agrupaciones de Estados y Carrocerias, así como las salidas de los Modelos Lineales Generalizados generados con R para las Frecuencias y Severidades de las tres coberturas restantes.

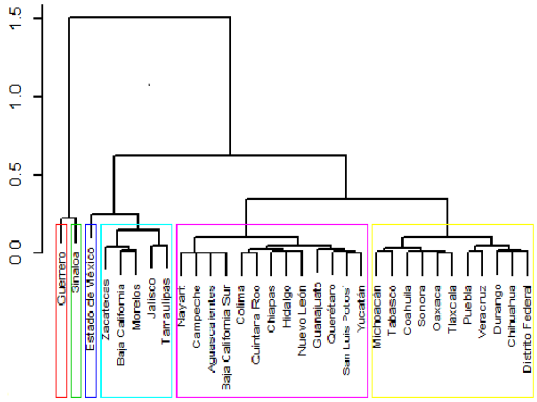
A1. Cobertura Robo Total

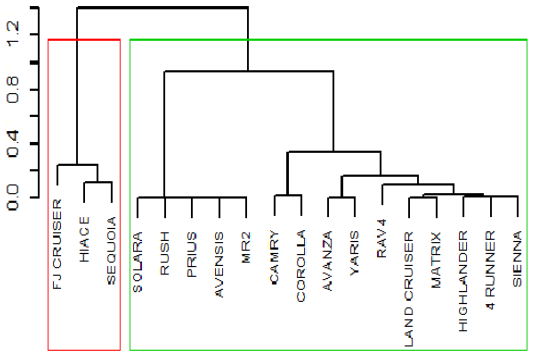
A2. Cobertura Responsabilidad Civil

A3. Cobertura Gastos Médicos

A1. Robo total

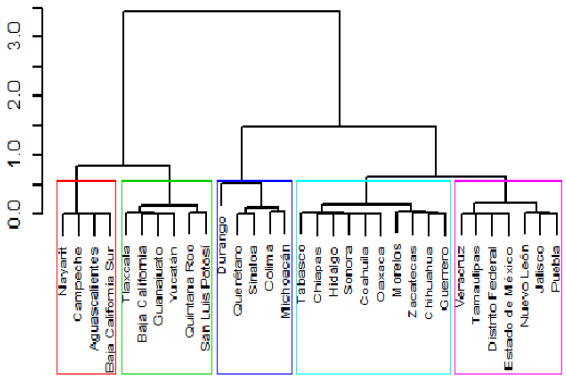
FRECUENCIA

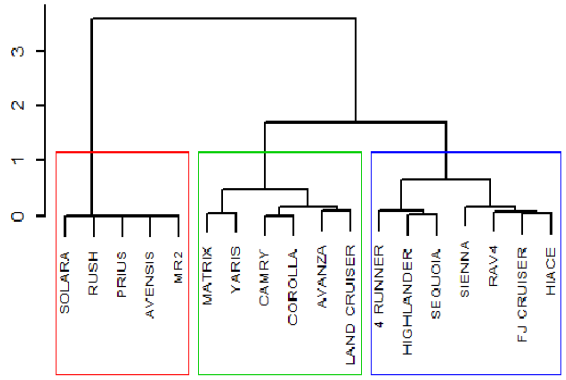





```
Call:
glm(formula = FRE$NS ~ FRE$GE + FRE$GC, family = poisson(link = log),
    offset = log(FRE$EXPUESTOS))

Deviance Residuals:
       1        2        3        4        5        6        7
       8        9       10       11
 0.39478  -0.18581  -0.14189  -0.31290  -0.03501   0.52871  -1.41432
0.56771   0.42241   0.97759   0.10174
      12
-2.15046

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.4838     0.1773 -42.205  < 2e-16 ***
FRE$GE2       2.0866     0.1927  10.828  < 2e-16 ***
FRE$GE3       1.2118     0.1951   6.212 5.23e-10 ***
FRE$GE4       2.5037     0.1965  12.742  < 2e-16 ***
FRE$GE5       3.0750     0.2946  10.437  < 2e-16 ***
FRE$GE6       3.3416     0.2007  16.652  < 2e-16 ***
FRE$GC2       0.8761     0.1414   6.197 5.76e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 552.9359  on 11  degrees of freedom
Residual deviance:   8.6807  on  5  degrees of freedom
AIC: 80.834

Number of Fisher Scoring iterations: 4
```

SEVERIDAD

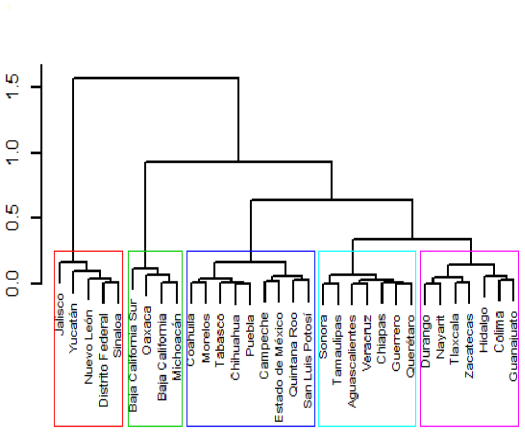


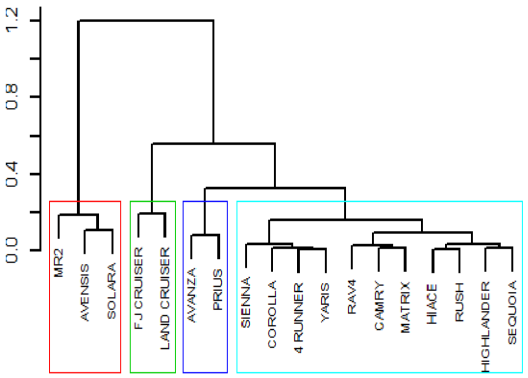



```
Call:
glm(formula = (SEV$SINOCU/SEV$NS) ~ SEV$GE + SEV$GC, family = Gamma(link = log))

Deviance Residuals:
       2        3        4        5        7        8        9       10
-0.12666   0.01498   0.01546   0.08817   0.11678  -0.01513  -0.01562  -0.09368

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.45180    0.09882 115.882 1.42e-06 ***
SEV$GE3      0.52502    0.12500   4.200  0.02463 *
SEV$GE4      1.18521    0.12500   9.482  0.00249 **
SEV$GE5      0.78896    0.12500   6.312  0.00804 **
SEV$GC2     -0.33691    0.08839  -3.812  0.03175 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.01562549)

    Null deviance: 1.739917  on 7  degrees of freedom
Residual deviance: 0.047164  on 3  degrees of freedom
  (7 observations deleted due to missingness)
AIC: 184.13

Number of Fisher Scoring iterations: 4
```

A2. Responsabilidad Civil

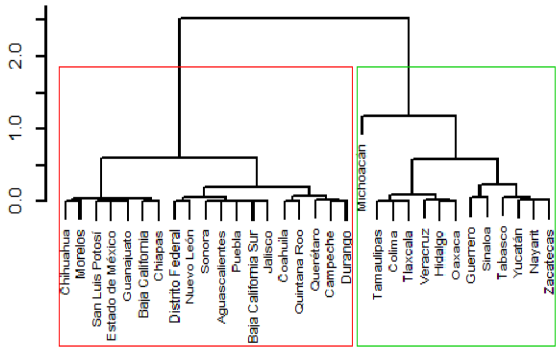
FRECUENCIA

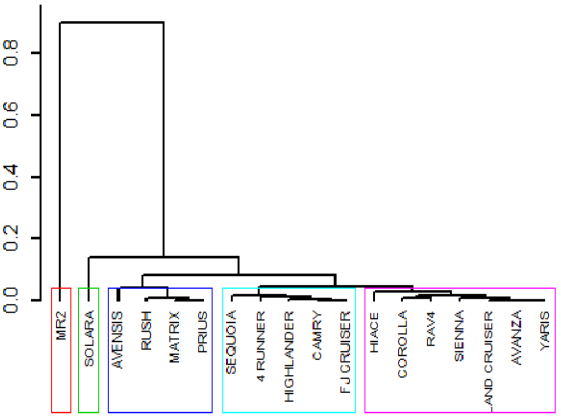





```
Call:
glm(formula = FRE$NS ~ FRE$GE + FRE$GC, family = poisson(link = log),
    offset = log(FRE$EXPUESTOS))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5057  -0.6974  -0.1859   0.5676   1.7520

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -2.99159    0.02327 -128.610  < 2e-16 ***
FRE$GE2     -0.51388    0.05051  -10.174  < 2e-16 ***
FRE$GE3      0.13784    0.02842    4.850 1.24e-06 ***
FRE$GE4     -0.18346    0.04025   -4.682 2.83e-06 ***
FRE$GE5      0.43127    0.02516   17.223  < 2e-16 ***
FRE$GC2     -0.23154    0.03332   -7.068 1.57e-12 ***
FRE$GC3     -0.61737    0.25673   -2.405   0.0162 *
FRE$GC4      0.31178    0.04813    6.478 9.30e-11 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1213.247  on 19  degrees of freedom
Residual deviance:   17.397  on 12  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 5
```

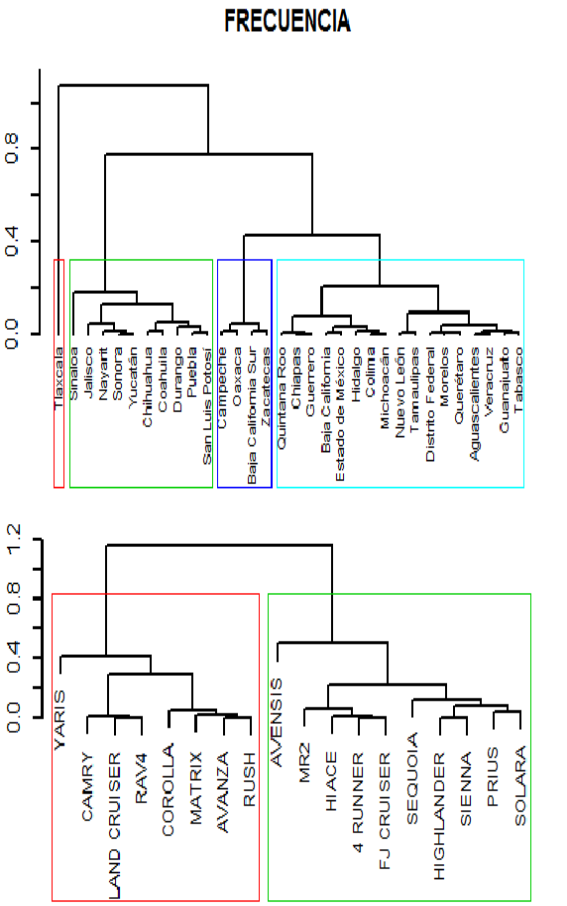
SEVERIDAD

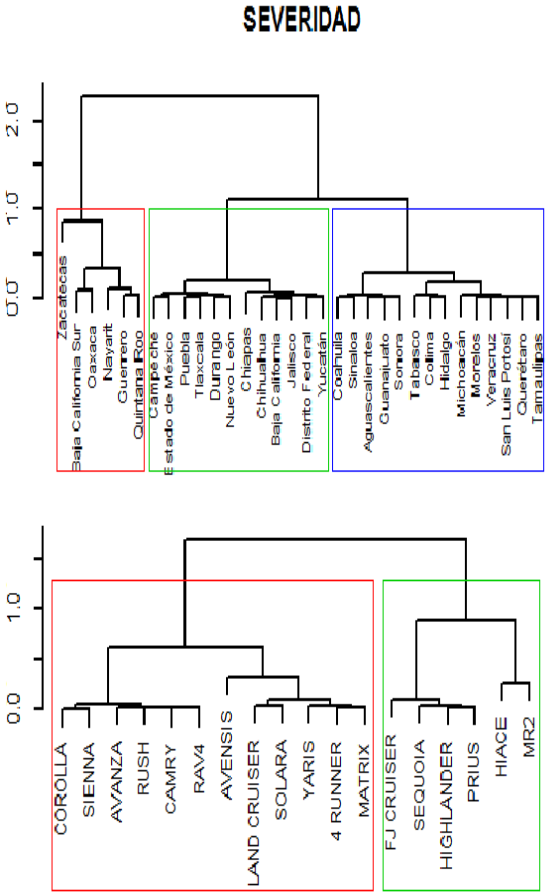





```
Call:
glm(formula = (SEV$SINOCU/SEV$NS) ~ SEV$GE + SEV$GC, family = Gamma(link = log))

Deviance Residuals:
       1         2         3         4         5         6         7         9
-0.02269   0.02235  -0.02323   0.02287   0.04489  -0.04628   0.00000   0.00000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.61168    0.04557 210.933 2.25e-05 ***
SEV$GE2      0.34164    0.04557   7.498  0.01733 *
SEV$GE3      0.10664    0.05581   1.911  0.04235 *
SEV$GC3      0.35053    0.05581   6.281  0.02442 *
SEV$GC4      1.24923    0.07205  17.339  0.00331 **
SEV$GC5      0.27698    0.07205   3.844  0.04149 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.003114585)

    Null deviance: 1.940680  on 7  degrees of freedom
Residual deviance: 0.006234  on 2  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 136.5

Number of Fisher Scoring iterations: 3
```

A3. Gastos Médicos






```
Call:
glm(formula = FRE$NS ~ FRE$GE + FRE$GC, family = poisson(link = log),
    offset = log(FRE$EXPUESTOS))

Deviance Residuals:
      1       2       3       4       5       6       7       8
-0.6106  -0.6050   0.8912  -0.6540   0.2454   0.2202  -0.3566   0.2925

Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -4.69048    0.04126 -113.675  < 2e-16 ***
FRE$GE2     -0.67333    0.14136   -4.763 1.91e-06 ***
FRE$GE3      0.36422    0.02992   12.172  < 2e-16 ***
FRE$GE4      1.14349    0.20942    5.460 4.75e-08 ***
FRE$GC2      0.67875    0.04239   16.011  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 509.5509  on 7  degrees of freedom
Residual deviance:   2.2823  on 3  degrees of freedom
AIC: 63.28

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = (SEV$SINOCU/SEV$NS) ~ SEV$GE + SEV$GC,
    family = Gamma(link = log))

Deviance Residuals:
      1       2       3       4       5       6
 0.02524   0.15790  -0.20561  -0.02567  -0.17655   0.18072

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.8175     0.2083  42.331 0.000558 ***
SEV$GE2      -0.6102     0.2551  -2.392 0.039218 *
SEV$GE3       0.8640     0.2551   3.387 0.047213 *
SEV$GC2       0.7732     0.2083   3.712 0.035523 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.0650815)

    Null deviance: 3.56652  on 5  degrees of freedom
Residual deviance: 0.13233  on 2  degrees of freedom
AIC: 115.5

Number of Fisher Scoring iterations: 5
```