UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Diciembre 2019

Un modelo jerárquico para determinar la demanda de seguros de vehículos a partir de la encuesta de hogares sobre demanda de inclusión financiera en seguros (Fasecolda 2018)

Hierarchical model to determine the demand for vehicles insurance

Eduar Velandia González[a]
eduarvelandia@usantotomas.edu.co

Deisy Yanira Camargo Galvis[b]
deisycamargo@usantotomas.edu.co

Wilmer Dario Pineda Rios[c]
wilmerpineda@usantotomas.edu.co

Oscar Estivens Velandia González[d]
ovelandia@fasecolda.com

Resumen

En los últimos años ha venido creciendo el número de hogares que son dueños o piensan en adquirir un vehículo lo que los hace potenciales clientes de un seguro de vehículo. El presente estudio tomó los resultados de la encuesta de hogares sobre demanda de inclusión financiera en seguros (Fasecolda 2018) y sobre esta se implementó un modelo jerárquico o modelo multinivel con el fin de aprovechar dicha información y poder determinar ¿cuáles son las características que tienen los hogares Colombianos a la hora de adquirir un seguro de vehículo? y a partir de estas características determinar la posibilidad de que adquiera un seguro los nuevos hogares que adquieran vehículos.

***Palabras clave*:** Encuesta de seguros, inclusión financiera, seguros, modelos jerárquicos, riesgo, modelos multinivel.

Abstract

In recent years the number of households that own or think about acquiring a vehicle has grown, which makes them potential customers of vehicle insurance. The present study took the results of the household survey on demand for financial inclusion in insurance (Fasecolda 2018) and on this a hierarchical model or multilevel model was implemented in order to take advantage of this information and be able to determine what are the characteristics that have Colombian homes when purchasing vehicle insurance? and from these characteristics determine the possibility that new homes that acquire vehicles acquire insurance.

***Keywords*:** Insurance survey, financial inclusion, insurance, hierarchies models, risk, multilevel models.

[a]Estudiante Estadística Universidad Santo Tomás
[b]Docente Estadística Universidad Santo Tomás
[c]Docente Estadística Universidad Santo Tomás
[d]Director de Actuaria de Fasecolda

1. Introducción

Las pólizas voluntarias de vehículos que se ofrecen actualmente en el mercado Colombiano están compuestas por tres coberturas: daños, hurtos y responsabilidad civil, estos son denominados seguros todo riesgo. Es necesario resaltar que estas pólizas en el país no son obligatorias y es voluntad de los propietarios de vehículos adquirir o no uno de estos seguros, a diferencia del Seguro Obligatorio de Accidentes de Tránsito, SOAT, el cual fue creado en la legislación Colombiana "mediante la Ley 33 de 1986, con el fin de garantizar los recursos que faciliten la atención integral para las víctimas de accidentes de tránsito" [1]. Cabe resaltar que este seguro obligatorio no tiene ninguna relación con el seguro de vehículos y su finalidad es de naturaleza diferente, el SOAT cubre lesiones o muerte del asegurado y lesiones personales a terceros pero no cubre daños a vehículo propio, daños a bienes de terceros ni hurto. Dado que el seguro de vehículos es una unión de varias coberturas que difieren entre si es necesario que tanto el asegurador como el asegurado las tengan en cuenta al momento de calcular la prima y administrar efectivamente el costo de los siniestros [1].

Para este trabajo se utilizaran datos publicados abiertamente por *Fasecolda* que es la federación de aseguradores Colombianos que agrupa a las compañías de seguros, reaseguros y a las sociedades de capitalización en Colombia. Esta entidad representa a la actividad del sector asegurador frente a las entidades de control y vigilancia así como la sociedad en general [1].

Según *Fasecolda* desde 2006 en Colombia, por medio del Plan Nacional de Desarrollo, se han establecido una serie de medidas para garantizar la oferta de servicios financieros, diseñar productos adecuados para los diferentes segmentos de la población y promover la educación económica y financiera, en aras de aumentar los índices de inclusión en el país. Diversos entes privados y públicos, como el Ministerio de Hacienda, la Superintendencia Financiera, Banca de las Oportunidades, el Banco de la República y los gremios del sector entre ellos Fasecolda, han tratado de aumentar el porcentaje de personas incluidas en el sistema financiero. *Fasecolda* en su libro Estudio de demanda de seguros 2018, expresa que ellos junto con la superintendencia financiera y la Banca de oportunidades (Bancoldex) crearon ¿ el primer estudio de demanda de inclusión financiera en seguros?, con el fin de conocer la situación de la inclusión de seguros en Colombia desde el punto de vista de la demanda. La recolección, procesamiento y análisis de la información fueron realizados por la unión temporal Cifras y Conceptos - Bioestadística, para la se tomó una muestra de 6520 hogares Colombianos, con representación a nivel nacional y regional, (para 7 regiones de Colombia).Los datos recogidos representan a 11,5 millones de hogares, de los cuales el 15,5 % corresponde a Antioquia; el 18,6 % a la región Atlántica, el 22,2 % a Bogotá; el 11,6 % a la región Central; el 16,2 % a la región Oriental; el 4,8 % a la región Pacífica; y el 11,1 % al Valle del Cauca [1]. Este trabajo de grado tiene como fin realizar un modelo jerárquico con la muestra ya creada y suministrada por la unión temporal Cifras y Conceptos - Bioestadística para Fasecolda (publicada en su página de internet) y con estos datos poder discriminar las características recurrentes de los hogares que poseen un seguro de vehículos.

2. Justificación

Dado el aumento en la demanda de vehículos en Colombia en los últimos años según cifras de Fenalco [6], es importante para las entidades que ofertan seguros en este ramo conocer las características que tienen los hogares que posiblemente pueden estar interesados en adquirir un seguro de vehículo.

Para una aseguradora es importante poder clasificar a sus potenciales clientes y saber cuales son las principales características de los hogares que han adquirido sus productos y servicios. Para ello es necesario estructurar, a que población le pueden ofrecer sus servicios con el fin de encontrar nuevos nichos de mercado en el cual haya mayor posibilidad de que los hogares se quieran asegurar.

Por esta razón el presente trabajo se realiza con el fin de generar mediante técnicas y herramientas estadísticas un modelo jerárquico de las características de los hogares colombianos aprovechando al

máximo la información obtenida en la encuesta realizada por Unión Temporal Cifras & Conceptos y Bioestadística para Banca de oportunidades y Fasecolda y con ello lograr encontrar y determinar el tipo de familias que en la encuesta reportan tener un seguro de vehículos.

3. Marco teórico

En Colombia en 2019 el numero de vehículos registrados ha presentado un crecimiento del 9 % en el primer semestre, con respecto al mismo periodo del 2018 según cifras del RUNT. En total los vehículos registrados son 408.020, de los cuales el 71 % son motocicletas (287.671), el 15 % de automóviles (63.310) y el 14 % restante (59.039) son otras clases de vehículos como camionetas, buses y camiones (véase [11]).

Junto al aumento en el parque automotor se ha elevado el índice de robos y siniestros de vehículos en Colombia, para el 2018 los datos arrojan 40.900 hurtos entre motos y autos, estas cifras según *Asopartes*. Por esta razón el sector asegurador esta interesado en encontrar las características que tienen los hogares Colombianos que afirman tener un seguro de autos, las cuales pueden ser diferentes de un departamento, a otro. Para así ofrecer la cobertura de seguros a más individuos que se encuentran expuestos a estos riesgos.

La siguiente tabla muestra el número de hurtos a vehículos entre 2018 y 2019 en Colombia:

Figura 1: **Evolución mensual del hurto de unidades del parque automotor entre 2018 - 2019 en Colombia**

Mes	Huerto de vehículos		Hurto de motocicletas		Total	
	2018	2019	2018	2019	2018	2019
Enero	606	664	2000	2121	2606	2785
Febrero	591	629	1900	1991	2491	2620
Marzo	597	619	1952	2026	2549	2645
Abril	619	634	2018	1900	2637	2534
Mayo	639	658	2064	1972	2703	2630
Junio	606	586	2050	1868	2656	2454
Julio	673	598	2066	1815	2739	2413
Agosto	670	505	2084	1380	2754	1885
Total	5001	4893	16134	15073	21135	19966

Fuente: ASOPARTE con base en datos de la dirección de investigación criminal e INTERPOL, Dijín, Policía Nacional, 2019

Como se puede ver en la figura el robo de vehículos creció en los primeros meses de 2019 lo que según cifras de Asopartes y la revista Dinero representa cerca de 685 millones de dólares en pérdidas [5], lo que lo hace un mercado en crecimiento para las aseguradoras.

A continuación se presenta el marco legal, demográfico y teórico de este documento.

3.1. Marco legal

Para el trabajo de grado se tuvo en cuenta la base de datos elaborada por la *Unión temporal Cifras & Conceptos y Bioestadística* para *Banca de Oportunidades* y *Fasecolda* [1]. Es una información de carácter público y por ende no aplica la Ley 1581 de 2012 que expidió el Régimen General de Protección de Datos, ya que esta base contiene información financiera, crediticia, comercial y de servicios, pero no posee datos personales.

3.2. Marco demográfico

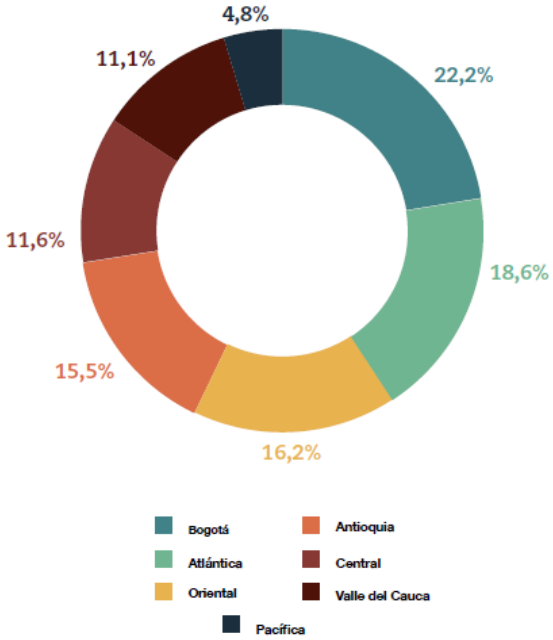
A continuación se presentan algunas características socio-demográficas del estudio de Demanda de seguros y que son importantes tener en cuenta a la hora de analizar la base y los resultados que se obtuvieron de ella. En el estudio de demanda de seguros 2018, la unión temporal tuvo como población objetivo los hogares colombianos y se tomo una muestra representativa de estos, en la cual fueron excluidos los siguientes departamentos: el archipiélago de San Andrés, Providencia, Santa Catalina, el Amazonas y la Orinoquia. Esto de debe a dos razones fundamentales, la primera porque estos departamentos tienen pocos habitantes y la segunda porque el costo de llegar a estos lugares es muy alto. Además, esta encuesta tomo tres niveles socio económicos en la población colombiana (bajo, medio y alto), donde se les indago acerca de aspectos característicos del hogar, que tanto conocían de los seguros o si tenían algún producto relacionado con los seguros.

3.2.1. Características de los hogares encuestados

El conjunto de datos es una muestra realizada a 6520 hogares Colombianos con 256 variables, las cuales tienen representación a nivel nacional, regional, (para 7 regiones de Colombia). Los datos recogidos representan a 11,4 millones de hogares de los cuales el 15,5 % corresponde a Antioquia, el 18,6 % a la región Atlántica, el 22,2 % a Bogotá, el 11,6 % a la región Central, el 16,2 % a la región Oriental, el 4,8 % a la región Pacífica, y el 11,1 % al Valle del Cauca [1].

La siguiente gráfica 2 muestra como se distribuyen los hogares de los departamentos encuestados:

Figura 2: *Distribución de los hogares que contestaron la encuesta por departamento*



3.2.2. Localización de los hogares

En el estudio, el 78.5 % de estos hogares están localizados en ciudades, el 12 % en municipios intermedios, el 6.2 % en municipios rurales y por último el 3.3 % en municipios rurales dispersos [1].

3.2.3. Nivel socio-económico

Por nivel socio-económico, el 62,8 % de los hogares corresponde al nivel bajo, el 28,9 % al nivel medio y el 8,3 % al nivel alto. El nivel bajo corresponde a los estratos 1 y 2, el medio al estrato 3 y el alto a los estratos 4, 5 y 6 [1].

3.2.4. Ingresos de los hogares

En el estudio el 18,6 % de los hogares declara ganar menos de quinientos mil pesos mensuales, el 36,4 % entre quinientos mil y un millón de pesos, el 29,6 % entre un millón y dos millones de pesos, el 5,7 % entre dos y tres millones de pesos y solo el 3,9 % declara ganar más de tres millones de pesos. El porcentaje restante de hogares (5,8 %) no quiso reportar sus ingresos [1].

En la siguiente tabla 3 se presentan los ingresos de las personas desagregadas por departamento:

Figura 3: *Rango de ingresos por departamento*

Rango de ingresos	Antioquia	Atlántica	Bogotá	Central	Oriental	Pacífica	Valle del Cauca
Menos de $500.000 pesos	24,4%	26,8%	9,0%	15,6%	17,4%	33,8%	13,9%
De $500.000 a menos de $1.000.000	31,3%	41,8%	28,9%	39,8%	36,5%	39,1%	44,6%
De $1.000.000 a menos de $1.500.000	16,5%	16,4%	22,5%	24,9%	16,6%	13,5%	23,2%
De $1.500.000 a menos de $2.000.000	8,1%	7,6%	15,9%	10,2%	8,5%	5,2%	10,2%
De $2.000.000 a menos de $3.000.000	4,4%	4,6%	9,5%	3,2%	6,9%	3,6%	3,7%
De $3.000.000 a menos de $4.000.000	0,9%	1,7%	3,9%	1,6%	2,5%	3,5%	1,8%
De $4.000.000 y más	0,6%	0,6%	3,1%	1,2%	1,8%	1,2%	2,1%
No sabe/ No responde	13,8%	0,5%	7,1%	3,5%	9,8%	0,1%	0,6%

3.2.5. Perfil de las personas que respondieron la encuesta

En el estudio, las personas que respondieron la encuesta solo podía ser el jefe del hogar o su pareja (Véase [1], pág. 14). Según el documento resumen, quienes respondieron fueron: 58,5 % mujeres y 41,5 % hombres.

Además, el 78,3 % se identificó como el jefe del hogar, y el 21,7 % como la pareja del jefe del hogar. Estos números sugieren que hay un alto porcentaje de hogares con jefatura femenina.

Por rangos de edad, quienes respondieron se clasifican así: entre 18 y 24 años el 6,3 %; entre 25 y 34 años: el 17,9 %; entre 35 y 44 años el 22,1 %; entre 45 y 54 años, el 21,5 %; entre 55 y 64 años el 18,1 %; más

de 65 años, el 14 %.

El 54,1 % de quienes respondieron la encuesta se declararon como trabajador, el 25,9 % dedicado a los oficios del hogar, el 6,5 % pensionado y el 6,4 % en busca de trabajo; y el porcentaje restante corresponde a otras actividades [1].

3.3. Modelo lineal generalizado

Según Rincón (2009) [12] un modelo lineal generalizado se utiliza para modelar un experimento en el cual la variable respuesta dependiente Y tiene una distribución que pertenece a la familia exponencial y está asociada a un conjunto de variables explicativas independientes $X_1, ..., X_p$,

$$g(\mathbb{E}(Y)) = X\beta.$$

En los modelos lineales generalizados se pueden distinguir tres componentes:

1. **Componente aleatoria**:

   Está representada por un conjunto de variables independientes Y_i con $i = 1, 2, ..., n$, cuya distribución para cada i pertenece a la familia exponencial, es decir, la función de densidad satisface [12]:

   $$f(y_i; \mu_i, \phi/q_i) = exp\left(\frac{q_i}{\phi}\left[y_i\theta_i - b(\theta_i) + c(y_i; \phi/q_i)\right]\right) \tag{1}$$

   Para algunas funciones $b(\cdot)$ y $c(\cdot)$ conocidas, y además

   a) $E(Y_i) = b'(\theta_i) = \mu_i$,

   b) $V(Y_i) = \frac{\phi}{q_i} b''(\theta_i) = \frac{\phi}{q_i} \cdot V(\mu_i)$.

2. **Componente sistemática**

   Está conformada por una matriz de variables aleatorias independientes $X_1, ..., X_p$ tal que

   $$\eta_i = \sum_{i=1}^{p} x_{ij}\beta_j \quad \text{equivalente a } \eta = X\beta$$

   la matriz X puede estar asociada a un modelo de regresión múltiple rango completo o incompleto [12].

3. **Función de Enlace**

   Es una función monótona, derivable que asocia o enlaza las componentes aleatoria y sistemática.

   $$g(\mu_i) = \eta_i$$

3.3.1. Modelo logístico

El modelo logístico es un caso en particular del modelo lineal generalizado descrito antes con las siguientes tres componentes [12]:

1. **Componente aleatoria**: Se asume que la variable respuesta U tiene distribución logística con parámetros μ y τ, la función de densidad de probabilidad está dada por:

$$f(u;\mu,\tau) = \frac{1}{\tau}\frac{e^{(u-\mu)/\tau}}{\left(1+e^{(u-\mu)/\tau}\right)^2}$$

Satisface $E(U) = \mu$ y $V(U) = \frac{\pi^2\tau^2}{3}$.

Reemplazando $\beta_1 = -\frac{\mu}{\tau}$ y $\beta_2 = \frac{1}{\tau}$, resulta la siguiente ecuación

$$f(u;\beta_1;\beta_2) = \beta_2\frac{e^{\beta_1+\beta_2 u}}{\left(1+e^{\beta_1+\beta_2 u}\right)^2}$$

2. **Función de enlace**: Para este modelo se utiliza como función de enlace la función logit, que está definida:

$$\eta_i = Logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) \tag{2}$$

3. **Predictor lineal** : Para el caso de una variable explicativa el predictor es $\eta_i = \beta_0 + \beta_1 x$ con la cual el modelo especificado es [12]:

$$Logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i \tag{3}$$

que se transforma en el modelo

$$E(Y_i) = \mu_i = m_i\frac{\exp^{(\beta_1+\beta_2 x_i)}}{1+\exp^{(\beta_1+\beta_2 x_i)}} \tag{4}$$

Algoritmo de estimación: es el proceso de estimación de los parámetros de un MLG y su método iterativo de *Newton-Raphson*, para la resolución de una ecuación basado en la aproximación de *Taylor*, para una función $f(x)$ de un punto de partida x_i,

$$f(x) = f(x_{i+1}) + (x - x_i)f'(x_i) = 0 \tag{5}$$

El proceso se repite como:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \tag{6}$$

Hasta alcanzar un valor suficientemente preciso. Según (Tjalling, (1995) [13] El método 5, y su extensión a la solución de sistemas de ecuaciones no lineales. La estimación de los coeficientes $\hat{\beta}_i$ del vector de parámetros del modelo, permite estimar en función $X_1, ..., X_p$, la probabilidad de que el suceso ocurra en función de π_i o el valor $E(Y_i)$.

3.3.2. Modelo Jerárquico

Según Gelman Andrew & Hill Jennifer (2006) [8] los modelos multinivel o jerárquicos, son extensiones de los modelos de regresión en los cuales la información está estructurada en grupos y los coeficientes de la regresión pueden variar por cada grupo que se estudia.

Los modelos jerárquicos pueden ser clasificados como: modelos de intersección variable, modelos de pendiente variable, o modelos de intersección y pendiente variables.

Suponga que la información puede ser agrupada en J grupos, si el modelo de regresión incluye un indicador para cada grupo en el intercepto, este modelo se denomina: *Modelo jerárquico de intercepto variable* y se nota cómo:

$$y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_i$$

donde cada $\alpha_{j[i]}$ es el intercepto del modelo de regresión en el j-ésimo grupo.

La segunda posibilidad con el modelo jerárquico es permitir que el vector de regresores varíe en cada grupo y mantener el intercepto constante, lo cual puede ser representado en la siguiente ecuación:

$$y_i = \alpha + \beta_{j[i]} \cdot x_i + \varepsilon_i$$

donde $\beta_{j[i]}$ representa una pendiente diferente para cada grupo de información, manteniendo el mismo intercepto.

Por ultimo, el modelo jerárquico permite dejar tanto el intercepto como el vector de regresores variables para cada grupo de la información, lo cual se puede representar en la siguiente ecuación:

$$y_i = \alpha_{j[i]} + \beta_{j[i]} \cdot x_i + \varepsilon_i$$

3.3.3. Modelo Lineal Generalizado Mixto

Según Correa & Salazar [2] en ocasiones es necesario modelar respuestas que no son necesariamente continuas pero que también se recolectan de manera longitudinal o por medio de mediciones repetidas, suponga la variable dependiente Y_i es el vector de datos observados. Además, asuma que la distribución de Y_i esta condicionada en b y son independientes, obtenidos de una distribución perteneciente a la familia exponencial. Entonces

$$f_{y_i|b}(y_i|b,\beta,\phi) = exp\left\{\frac{y_i\eta_i + c(\eta_i)}{a(\phi)} + d(y_i,\phi)\right\} \tag{7}$$

$$b \sim f_b(b|D)$$

El modelo lineal generalizado mixto *MLMG* está formulado como:

$$\eta_i = X_i^{'}\beta + z_i^{'}b$$

con $X_i^{'}$ la i-ésima fila de X y con $z_i^{'}$ la i-ésima fila de Z. Las distribuciones normal, binomial y Poisson son miembros de esta familia y cada una de ellas tiene asociado un importante modelo de regresión. Además, la *Vesorimilitud* esta dada por esta función no puede evaluarse en forma cerrada y requiere de métodos numéricos. El método de la cuadratura de Gauss pueden ser una opción adecuada para esta evaluación.

$$L\left(\beta,\phi,D|Y\right) = \int \prod_{i=1}^{n} f_{y_i|b}(y_i|b,\beta,\phi) f_b(b|D) db \tag{8}$$

Inicialmente se presenta la aproximación el modelo lineal mixto tradicional con la siguiente estructura

$$\begin{aligned} Y &= \eta + \epsilon \\ &= X\beta + Zb + \epsilon \end{aligned} \tag{9}$$

donde $\epsilon \sim N(0, \sigma^2 D)$, con D conocida. El vector de respuesta media η depende de una componente fija $X\beta$, con X la matriz de diseño con constantes conocidas y β el vector de parámetros poblacionales desconocidos. Además, asuma que Zb puede ser particionado como:

$$Z = (Z_1, \ldots, Z_k)$$
$$b^{'} = (b_1^{'}, \ldots, b_k^{'})$$

donde b_j tiene v_j componentes y $b_j \sim N(0, \sigma_j^2 A_j)$ y además es independiente de las otras componentes b. Sea $\sigma_j^2 = \sigma^2 \theta_j$, entonces

$$A = \begin{bmatrix} \theta_1 A_1 & 0 & \cdots & 0 \\ 0 & \theta_2 A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_k A_k \end{bmatrix} \tag{10}$$

El procedimiento BLUP consiste en la maximización de la suma de dos componentes de log-verosimilitud. Sea

1. Sea l_1 la log-verosimilitud para Y dado b condicionalmente fijo,
2. sea l_2 la log-verosimilitud para b y
3. $l = l_1 + l_2$

Así, l representa la *log-verosimilitud* basada en la distribución conjunta de Y y b.

El procedimiento BLUP selecciona estimaciones o predicciones de β, b, σ^2 y los parámetros de A que maximicen l. Resolviendo el sistema de ecuaciones se obtiene

$$\begin{bmatrix} X^{'} D^{-1} X & X^{'} D^{-1} Z \\ Z^{'} D^{-1} X & Z^{'} D^{-1} Z + A^{-1} \end{bmatrix} \cdot \begin{bmatrix} \tilde{\beta} \\ \tilde{b} \end{bmatrix} = \begin{bmatrix} X^{'} D^{-1} Y \\ Z^{'} D^{-1} Y \end{bmatrix} \tag{11}$$

3.4. Criterio de información Akaike

El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos, es una compensación entre la bondad de ajuste del modelo y la complejidad del modelo. Se basa en la entropía de información: se ofrece una estimación relativa de la información pérdida cuando se utiliza un modelo determinado para representar el proceso que genera los datos.

El valor de AIC no proporciona una prueba de hipótesis, es decir no existe una *hipótesis nula* que debe ser *rechazada* o *no rechazada*, es decir, no se puede decir nada acerca de la calidad del modelo en un sentido absoluto.

3.5. Selección de variables

En muchas situaciones se dispone de un conjunto de posibles variables regresoras, sin embargo, una pregunta que puede surgir es *¿todas las variables deben entrar en el modelo?*, en caso negativo, existen métodos para seleccionar las variables dentro de un modelo de regresión.

Algunos de estos métodos son:

1. *Eliminación progresiva* (*backward stepwise*): parte del modelo de regresión con todas las variables regresoras y en cada etapa se elimina la variable menos influyente según el contraste individual de la T o de la F hasta una cierta regla de parada. El procedimiento de eliminación progresiva tiene los inconvenientes de necesitar una alta capacidad de cálculo computacional, si se cuenta con un conjunto de datos grande y una cantidad considerable de variables, adicional puede llevar a problemas de multicolinealidad si las variables están relacionadas. Tiene la ventaja de no eliminar variables significativas.

2. *Introducción progresiva* ($fordward\ stepwise$): Este algoritmo funciona de forma inversa que el anterior, parte del modelo sin ninguna variable represora y en cada etapa se introduce la más significativa. El procedimiento de introducción progresiva tiene la ventaja respecto al anterior de necesitar menos capacidad de cálculo computacional, pero presenta dos graves inconvenientes, el primero, que pueden aparecer errores de especificación porque las variables introducidas permanecen en el modelo, aunque el algoritmo en pasos sucesivos introduzca nuevas variables que aportan la información de las primeras. Este algoritmo también falla si el contraste conjunto es significativo pero los individuales no lo son, ya que no introduce variables regresoras.

3. *Regresión paso a paso* (*stepwise regression*): Este método es una combinación de los procedimientos anteriores, comienza como el de introducción progresiva, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer. Termina el algoritmo cuando ninguna variable entra o sale del modelo. El algoritmo paso a paso tiene las ventajas del algoritmo de introducción progresiva, pero lo mejora al no mantener fijas en el modelo las variables que ya entraron en una etapa, evitando de esta forma problemas de multicolinealidad. En la práctica, es un algoritmo bastante utilizado que proporciona resultados razonables cuando se tiene un número grande de variables regresoras.

Para este trabajo se utilizó un modelo Jerárquico mixto, con cual se modelaron varios escenarios de los cuales se escogió el que presentaba menor AIC. Sobre este modelo se graficaron la razón de ODDS y se presentarán las conclusiones. El paso a paso del método utilizado se presenta en la siguiente sección.

4. Metodología

Este proyecto es un estudio con enfoque mixto, transversal retrospectivo y a su vez descriptivo etnográfico. Se indaga sobre la información de los hogares que respondieron a la encuesta de demanda de seguros 2018 basados en el instrumento creado por la unión temporal Cifras y Conceptos & Bioestadística, para poder encontrar las características de los hogares que manifiestan poseer un seguro de vehículo y de esta forma identificar posibles clientes o posibles nichos de negocio no explorados por las compañías de seguros colombianas.

El conjunto de datos cuenta con una muestra realizada a 6520 hogares Colombianos con 256 variables las cuales tienen representación a nivel nacional, regional (7 regiones de Colombia). Como se explico anteriormente los datos recogidos representan a 11,4 millones de hogares los cuales se distribuyen en las regiones de: Antioquia, Atlántica, Bogotá, región Central, región Oriental, Pacífica y Valle del Cauca [1].

En primera instancia se hizo un análisis descriptivo de las variables y se formuló un modelo lineal generalizado *Logit*, el cual buscaba encontrar las covariables que son significativas en función de la variable respuesta binaria donde cero (0) es "no es poseedor de un seguro de vehículo" y uno (1) es "posee un seguro de vehículo".

En la siguiente tabla 1 se muestra el conjunto inicial de variables:

Tabla 1: **Descripción y estructura de la información utilizada.**

Variable	Tipo	Descripción
Edad	numérico	edad de quien responde la encuesta
Nombre Departamento	nominal	Departamento
Mujer Jefe de Hogar	nominal	Mujer Jefe de Hogar (si o no)
En este hogar son propietarios de vehículos	nominal	(si o no)
En este hogar son propietarios de motos	nominal	(si o no)
Respetan las señales de tránsito	nominal	(si o no)
Posee SOAT	nominal	(si o no)
Tienen seguro de vehículo (Carro o moto)	nominal	(si o no)
Seguro de Vida	nominal	(si o no)
Seguro exequial	nominal	(si o no)
Seguro de accidentes escolares	nominal	(si o no)
EPS - POS	nominal	(si o no)
Plan complementario al POS o medicina pre pagada	nominal	(si o no)
Seguro educativo	nominal	(si o no)
Seguro de incendio y terremoto - Hogar	nominal	(si o no)
Seguro de incendio y terremoto - Negocio	nominal	(si o no)
Seguro de desempleo	nominal	(si o no)
ARL	nominal	(si o no)
Seguro agropecuario	nominal	(si o no)
Responsabilidad civil extracontractual	nominal	(si o no)

Fuente: elaboración propia.

Como segunda fase se tiene en cuenta el factor de expansión y se propone un modelo donde estén todas estas variables para observar cual es el aporte de cada una de ellas, como se menciono anteriormente se tendrá en cuenta el menor valor de AIC. Para comenzar se arranco con un modelo de 19 variables el cual arroja un AIC de 3921.252; a continuación se presenta tabla resumen

```
Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -4.963174   0.243586 -20.375  < 2e-16 ***
A3..Edad                                          -0.008402   0.003064  -2.742  0.00611 **
C2...Los.miembros.de.este.hogar.son.propietarios.de.vehículos.1  2.316444   0.109191  21.215  < 2e-16 ***
C5...Los.miembros.de.este.hogar.son.propietarios.de.motos.1      0.300351   0.097999   3.065  0.00218 **
F2.2.Respetan.las.señales.de.tránsito1            0.352872   0.176418   2.000  0.04548 *
G1.13..SOAT1                                      1.830331   0.142742  12.823  < 2e-16 ***
Mujer.Jefe..Hogar1                               -0.249319   0.095141  -2.621  0.00878 **
G1.1..Seguro.de.Vida1                             0.482412   0.093719   5.147 2.64e-07 ***
G1.2.Seguro.exequial1                            -0.078146   0.092281  -0.847  0.39710
G1.3.Seguro.de.accidentes.escolares1              0.091621   0.103549   0.885  0.37626
G1.5.EPS...POS1                                   0.101618   0.121106   0.839  0.40142
G1.6.Plan.complementario.al.POS.o.medicina.prepagada1  0.601902   0.115805   5.198 2.02e-07 ***
G1.7..Seguro.educativo1                           0.623099   0.191835   3.248  0.00116 **
G1.8..Seguro.de.incendio.y.terremoto...Hogar1     0.576608   0.177401   3.250  0.00115 **
G1.9..Seguro.de.incendio.y.terremoto...Negocio1   0.301273   0.401171   0.751  0.45266
G1.10..Seguro.de.desempleo1                       0.368225   0.191461   1.923  0.05445 .
G1.11..ARL1                                       0.006804   0.092620   0.073  0.94143
G1.12..Seguro.agropecuario1                       3.813387   0.720652   5.292 1.21e-07 ***
G1.15..Responsabilidad.civil.extracontractual1    1.565837   0.223362   7.010 2.38e-12 ***
```

Figura 4: *Resumen de Modelo 1*

En la tabla anterior se observa que existen variables que no son significativas como: la tenencia de un seguro exequial, la tenencia de un seguro de accidentes escolares, tener EPS o POS, tener un seguro de incendio y terremoto para negocio y por ultimo si la persona tiene ARL. Por lo cual se realiza un

proceso de selección de variables por el método iterativo de stepwase y de esta forma encontrar la mejor combinación de variables que genere el menor AIC.

```
Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -4.887230   0.230244 -21.226  < 2e-16 ***
A3..Edad                                          -0.008789   0.002985  -2.944 0.003239 **
C2...Los.miembros.de.este.hogar.son.propietarios.de.vehículos.1  2.323807   0.108766  21.365  < 2e-16 ***
C5...Los.miembros.de.este.hogar.son.propietarios.de.motos.1      0.306484   0.097564   3.141 0.001682 **
F2.2.Respetan.las.señales.de.tránsito1            0.334847   0.175393   1.909 0.056246 .
G1.13..SOAT1                                      1.840379   0.140826  13.068  < 2e-16 ***
Mujer.Jefe..Hogar1                                -0.244530   0.094970  -2.575 0.010030 *
G1.1..Seguro.de.Vida1                             0.475657   0.091344   5.207 1.92e-07 ***
G1.6.Plan.complementario.al.POS.o.medicina.prepagada1  0.597093   0.114214   5.228 1.71e-07 ***
G1.7..Seguro.educativo1                           0.655993   0.188961   3.472 0.000517 ***
G1.8..Seguro.de.incendio.y.terremoto...Hogar1     0.602524   0.171517   3.513 0.000443 ***
G1.10..Seguro.de.desempleo1                       0.399997   0.188106   2.126 0.033466 *
G1.12..Seguro.agropecuario1                       3.839135   0.726582   5.284 1.27e-07 ***
G1.15..Responsabilidad.civil.extracontractual1    1.561729   0.223513   6.987 2.80e-12 ***
```

Figura 5: *Resumen de Datos Modelo 2*

En esta tabla se observa que todas las variables son significativas y aportan al modelo, este modelo mejora con respecto al anterior a partir del criterio de Akaike con una medida de AIC de 3913.914. En esta fase previa se encontró el mejor conjunto de variables, lo que se hace a continuación es crear el modelo jerárquico mixto a partir de ella, con la condición de pendiente e intercepto variable y observar que impacto tiene ajustar este modelo al conjunto de variables escogidas.

Al calcular un modelo jerárquico con pendiente e intercepto variable, donde la variable cambiante por cada departamento es "mujer jefe cabeza de hogar" se obtuvo el siguiente modelo resumido en la siguiente tabla,

```
Fixed effects:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -5.420896   0.295775 -18.328  < 2e-16 ***
A3..Edad                                          -0.005245   0.003116  -1.683 0.092317 .
C2...Los.miembros.de.este.hogar.son.propietarios.de.vehículos.1  2.355051   0.113379  20.772  < 2e-16 ***
C5...Los.miembros.de.este.hogar.son.propietarios.de.motos.1      0.475781   0.107585   4.422 9.76e-06 ***
F2.2.Respetan.las.señales.de.tránsito1            0.145096   0.182669   0.794 0.427015
G1.13..SOAT1                                      1.909613   0.144838  13.184  < 2e-16 ***
Mujer.Jefe..Hogar1                                -0.048814   0.181965  -0.268 0.788499
G1.1..Seguro.de.Vida1                             0.429737   0.095467   4.501 6.75e-06 ***
G1.6.Plan.complementario.al.POS.o.medicina.prepagada1  0.599651   0.118697   5.052 4.37e-07 ***
G1.7..Seguro.educativo1                           0.703514   0.196563   3.579 0.000345 ***
G1.8..Seguro.de.incendio.y.terremoto...Hogar1     0.640659   0.176733   3.625 0.000289 ***
G1.10..Seguro.de.desempleo1                       0.390940   0.195556   1.999 0.045596 *
G1.12..Seguro.agropecuario1                       3.587394   0.688858   5.208 1.91e-07 ***
G1.15..Responsabilidad.civil.extracontractual1    1.598380   0.231301   6.910 4.83e-12 ***
```

Figura 6: *Resumen de Datos Modelo Jerárquico*

Se puede observar que las variables siguen siendo significativas y se reduce la medida del criterio de Akaike con AIC a 3753.869.

En la siguiente sección se presentan los resultados que arroja el modelo y las conclusiones a las que se puede llegar.

5. Gráficos y Resultados

En el siguiente gráfico se muestra cuales son los departamentos en los que hay mayor éxito a la hora de poseer un seguro de vehículos. De estos, se resalta el Valle del Cauca en donde un hogar que cuente: con una mujer cabeza de hogar y el resto de variables permanezcan constantes, la posibilidad de que ese

hogar adquiera un seguro es dos veces mayor que el chance de poseerlo en un hogar donde una mujer no sea jefe de hogar.

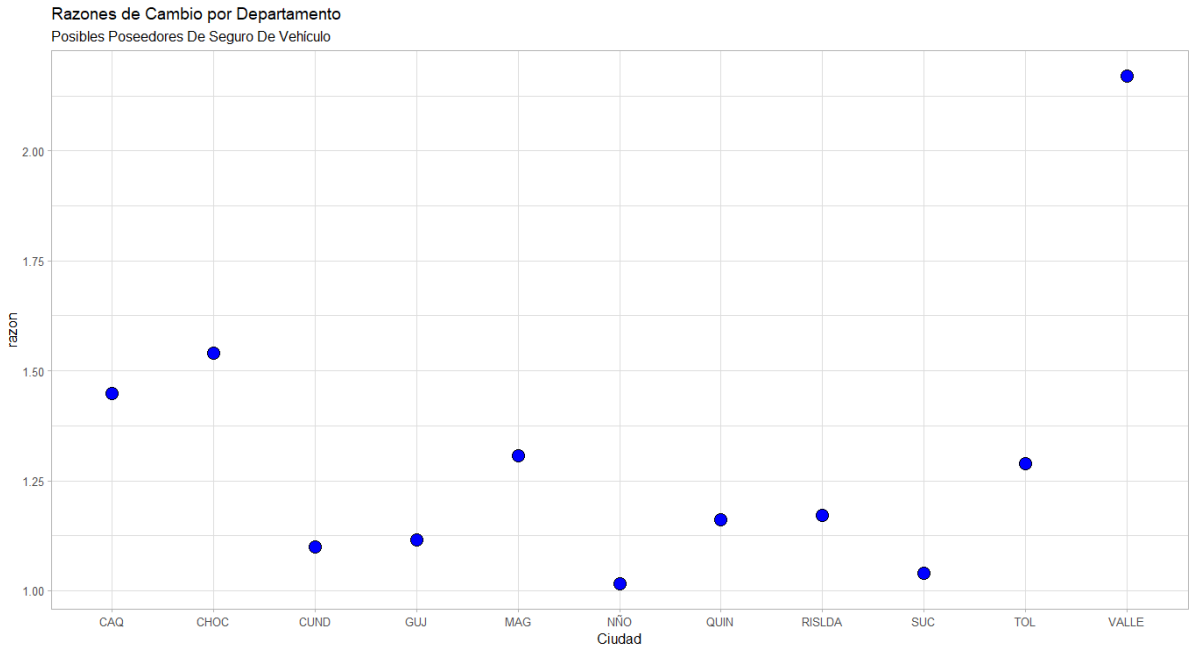


Figura 7: *Grafico Razon*

Por otro lado, en la siguiente gráfica se presentan los hogares donde es menor el chance dado que se mantienen las mismas condiciones. Es de resaltar que el departamento de Antioquia es el que tiene menos chance de adquirir un seguro de vehículos, cuando el hogar cuenta con una mujer como cabeza de hogar.

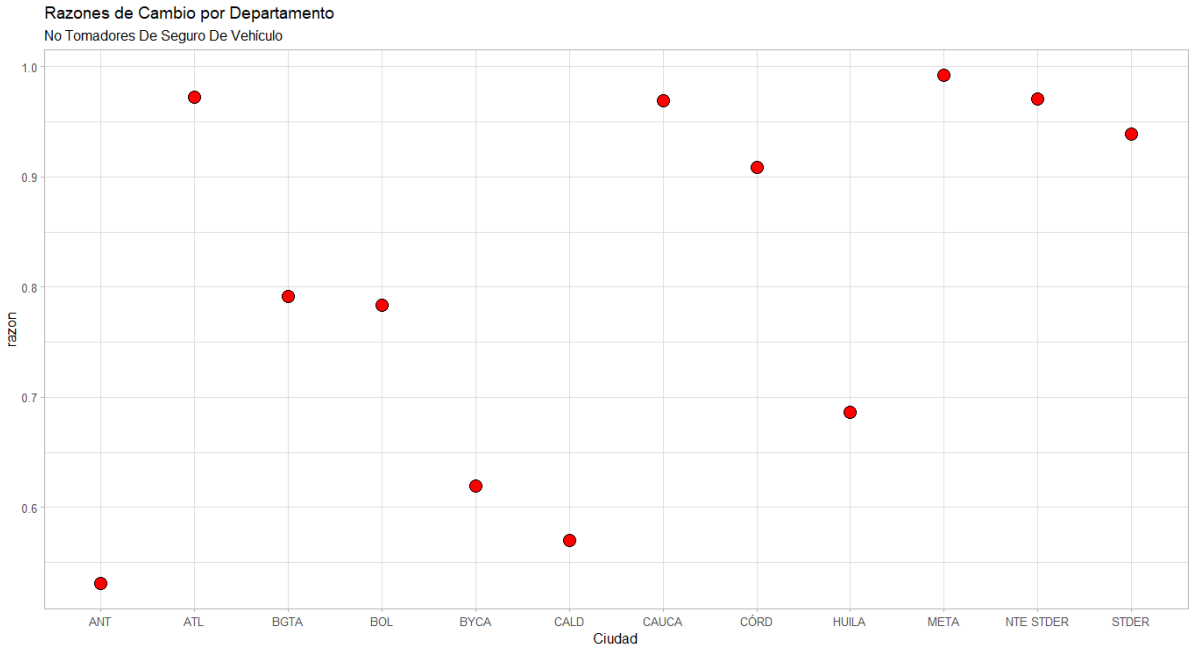


Figura 8: *Gráfico Razón de ODDS*

6. Conclusiones

1. Se encontró que poseer otros seguros es influyente a la hora de establecer el chance de que un hogar posea un seguro de vehículo. Los seguros más influyentes en el modelo fueron: agropecuario, responsabilidad civil, SOAT y seguro Educativo.

2. Al generar un modelo con intercepto y pendiente variable se pudo desagregar la información por departamento lo que le permite a las aseguradoras focalizar los esfuerzos en función de los hallazgos por departamento.

3. La variable "mujer jefe cabeza de hogar" presenta razón de ODDS por encima de uno en los departamentos de Valle del Cauca, Choco, Caqueta, Magdalena entre otros, lo que implica un mayor chance de tener un seguro de vehículos en hogares con estas características.

4. La variable "mujer jefe cabeza de hogar" presenta razón de ODDS por debajo de uno en los departamentos de Antioquia, Caldas, Boyacá, Huila, entre otros; lo que implica un menor chance de tener un seguro de vehículos en hogares con estas características.

5. En el modelo jerárquico se detecto que la edad tiene un impacto importante e inverso. Entre más edad, menor es el chance de poseer un seguro de vehículos.

Referencias

[1] Banca de las Oportunidades, Fasecolda y la Superintendencia Financiera de Colombia (2018) *Estudio de demanda de seguros en Colombia 2018*

[2] Correa Morales, Juan Carlos& Salazar Uribe, Juan carlos (2016) *Introducción a los modelos mixtos*,Universidad Nacional de Colombia, ISBN 978-958-775-953-2

[3] De la Pava Roys Nasser Stephan (2019) Trabajo de Grado Universidad Santo Tomás *Detección de fraude de energía: comparación entre un modelo de lógica difusa y un MLG logit.*

[4] Desjardins, D., Dionne, G. y Pinquet, J. (2001) *Esquemas de calificación de experiencia para flotas de vehículos.* Boletín ASTIN, 31 (1), 81-105. doi: 10.2143 / AST.31.1.995

[5] Dinero. *En 2018 se disparó el hurto de vehículos y motos*, Revista Dinero, 31 Enero. 2019 (https://www.dinero.com/economia/articulo/cuantos-vehiculos-se-robaron-en-colombia-en-2018/266636)

[6] FENALCO & ANDI *Informe del sector automotor a Agosto de 2019*, (2019).

[7] Frees Edward W. y Valdez Emiliano A. (2008) *Modelado de reclamos de seguros jerárquicos* Journal of the American Statistical Association, 103: 484, 1457-1469, DOI: 10.1198 / 016214508000000823

[8] Gelman Andrew & Hill Jennifer (2006) *Data Analysis Using Regression and Multilevel Hierarchical Models* First Edition, ISBN 978-1-78216-214-8

[9] González Martínez Edwin Fernando (2018) Trabajo de Grado Universidad Santo Tomás *Detección de Fraude en Tarjetas de Crédito Mediante Técnicas de Minería de Datos.*

[10] Gutierrez Andres (2015) *Estrategias de muestreo Diseño de encuestas y estimación de parámetros* Second Edition,ISBN: 978-958-631-608-8.

[11] Mintransporte & RUNT (2019) *Boletín de Prensa 005 de 2019*

[12] Rincón Suárez Luis Francisco (2009) *Curso Básico de Modelos Lineales.* Universidad Santo Tomás

[13] Tjalling J. Ypma, *Historical development of the Newton Raphson method*, SIAM Review 37 (4), 531-551, 1995. doi:10.1137 / 1037125

[14] Torgo Luis (2011) *Data Mining with R Learning With Case Studies* Chapman & Hall / CRC, ISBN 9781439810187