© Revista de Matemática: Teoría y Aplicaciones 2021 **28**(1) : 105–124
cimpa – ucr issn: 1409-2433 (Print), 2215-3373 (Online)
doi: https://doi.org/10.15517/rmta.v28i1.39030

MODELACIÓN PREDICTIVA DE SINIESTROS EN SEGUROS DE NO VIDA

PREDICTIVE MODELLING OF LOSSES IN NON-LIFE INSURANCE

Ana Rosa Sandí-Corrales*

Received: 20/Feb/2020; Revised: 16/Oct/2020;
Accepted: 30/Oct/2020

*Universidad de Costa Rica, Escuela de Matemática, San José, Costa Rica. E-Mail: ana.sandicorrales@ucr.ac.cr

Resumen

Se analizó un seguro de accidentes y salud que tiene primas diferenciadas para grupos de riesgo homogéneos. La estimación de dichas primas en ocasiones anteriores fue de tipo univariado, que tiene la limitante de que cuando hay grupos de riesgo con pocas observaciones los resultados son muy volátiles y omiten la información que podrían aportar variables predictoras. Por lo que se optó por estimar los siniestros esperados (que son insumo del cálculo de primas) con tres modelos multivariados: lineales ordinarios, aditivos y lineales mixtos. Se utilizaron varios con el fin de comparar su capacidad de pronóstico. El desempeño fue aceptable tanto dentro de la muestra de ajuste como de prueba en el caso de los modelos lineal ordinario y aditivo con una diferencia porcentual de alrededor del 1% con respecto a los datos reales. El lineal mixto no pudo hacer pronósticos para combinaciones de predictores no observados en los datos de ajuste.

Palabras clave: seguros; tarificación; modelación predictiva; modelos lineales; modelos aditivos; modelos mixtos; paquete estadístico R.

Abstract

Accident and health insurance with differentiated premiums for homogeneous risk groups was analyzed. The estimation of these premiums on previous opportunities was in univariate form, which has the limitation that when there are risk groups with few observations, the results are very volatile and omit the information that could provide predictive variables. Therefore, it was decided to estimate the expected claims (which are an input in the premium calculation) with three multivariate models: ordinary linear, additive and mixed linear. Several were used in order to compare their forecasting capability. Performance was acceptable within both the fit and test samples in the case of ordinary linear and additive models with a difference of about 1% from the real data. Linear mixed could not make predictions for combinations of predictors not observed in the fit data.

Keywords: insurance; pricing; predictive modelling; linear models; additive models; mixed models; software R.

Mathematics Subject Classification: 62P05.

1 Introducción

En los seguros de no vida existe la particularidad de que no es posible conocer de antemano el monto de la indemnización (severidad), porque depende de la pérdida económica que genere el evento, y al mismo tiempo, se desconoce el momento en que este ocurrirá (frecuencia), por lo que hay una doble incertidumbre. Otro elemento a considerar es que las compañías aseguradoras cobran primas diferenciadas por grupos de acuerdo con su exposición al riesgo.

En la actualidad existe una amplia gama de modelos estadísticos que se han implementado para proyectar los siniestros esperados como lo son los modelos lineales generalizados, para ello se puede ver el capítulo 5 de [6] escrito por Curtis Gary Dean, o el capítulo 2 de [7] escrito por Dan Tevet; lineales mixtos (Antonio y Zhang en [6]); no lineales (Antonio y Zhang en [6]); bayesianos (Nieto-Barajas y de Alba en [6]); y no paramétricos.

En este proyecto se utilizó información estadística de una cartera de asegurados del ramo de accidentes y salud durante el periodo 2008-2012 y se evaluó la calidad del pronóstico en el intervalo 2013-2015 para los modelos citados.

Adicionalmente, se buscó con este trabajo valorar la capacidad de los modelos escogidos para estimar siniestros en subgrupos con poca información y si las variables predictoras son relevantes para ser consideradas en el proceso de suscripción.

En el análisis descriptivo y estimación de modelos se utilizó el software estadístico R. La lista de paquetes empleados se menciona a lo largo del documento.

2 Modelos multivariados

Dado que existen modelos paramétricos, semi y no paramétricos, se escogieron representantes de cada tipo para valorar su calidad de pronóstico. Los primeros son más simples de estimar y de interpretar, ya que se expresan como una fórmula cerrada. Los no paramétricos, por su parte, son más flexibles, tratan de buscar el modelo que mejor se ajusta, impidiendo errores de escogencia y sirven si hay pocos datos [5]. De seguido se exponen brevemente los modelos seleccionados para su estudio.

2.1 Modelos lineales ordinarios (OLS)

El modelo estima el valor que tomará la variable dependiente $\overrightarrow{y} \in \mathbb{R}^n$ a partir de $p-1$ variables explicativas o predictivas $X_1, ..., X_{p-1}$, las cuales también pertenecen a $\mathbb{R}^n$. Con estas variables se construye la matriz de diseño X de tamaño n x p, cuya primer columna es un vector de unos de largo n que representa la intersección y las siguientes columnas son las $p-1$ variables predictivas. Se interpreta que para cada uno de los predictores (columnas) se cuenta con n observaciones.

Se establece una relación lineal con los predictores, con expresión $\overrightarrow{y} = X \cdot \overrightarrow{\beta} + \overrightarrow{\varepsilon}$, donde $\overrightarrow{\beta}$ se conoce como el vector de coeficientes de regresión y $\overrightarrow{\varepsilon}$ el vector de errores de estimación. Se anota que $\overrightarrow{\beta} \in \mathbb{R}^p$ y $\overrightarrow{\varepsilon} \in \mathbb{R}^n$ [4].

En la regresión lineal se hace el supuesto de que los errores son aditivos, no correlacionados y con distribución $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d. con varianza constante (homocedasticidad). Los coeficientes $\overrightarrow{\beta}$ se obtienen por el criterio de mínimos cuadrados mediante la reducción del error. Si se asume que $X^T \cdot X$ es invertible, o al menos existe la inversa generalizada, la solución es:

$$\overrightarrow{\hat{\beta}} = \left(X^T \cdot X\right)^{-1} X^T \cdot \overrightarrow{y} \tag{1}$$

Cuando los errores se distribuyen normalmente y son no correlacionados, el teorema de Gauss-Markov permite inferir que $\overrightarrow{\hat{\beta}}$ es el mejor estimador lineal insesgado (BLUE por sus siglas en inglés) teniendo varianza mínima [4].

En [8] se menciona que la estimación por mínimos cuadrados descubierto por Gauss coincide con la de máxima verosimilitud cuando está presente el supuesto de distribución normal de los errores. Cuando la hipótesis no se cumple, la verosimilitud produce mejores estimadores.

2.2 Modelos aditivos (AM)

Se ubican en punto intermedio entre los métodos paramétricos y no paramétricos. Las transformaciones a los predictores se encuentran simultáneamente sin hacer supuestos sobre distribuciones, si no utilizando suavizadores, que a continuación se explican.

Suavizadores

Un suavizador "scatterplot" $\overrightarrow{y} - \overrightarrow{x}$ entre la respuesta $\overrightarrow{y}$ y el predictor $\overrightarrow{x}$ se define como una función $S\left(x_0 \mid \overrightarrow{x}, \overrightarrow{y}\right)$ en la que para cada punto x_0 estima la dependencia de $\overrightarrow{y}$ sobre $\overrightarrow{x}$. El dominio de la función es el rango de $\overrightarrow{x}$.

Si se opta por un suavizador lineal sucede que si se escogen pesos $s\left(i, x_0, \overrightarrow{x}\right)$ adecuados, donde i representa la i-ésima entrada del predictor $\overrightarrow{x}$, el suavizador lineal toma la forma:

$$S\left(x_0 \mid \overrightarrow{x}, \overrightarrow{y}\right) = \sum_{i=1}^{n} s\left(i, x_0, \overrightarrow{x}\right) \cdot y_i. \tag{2}$$

Existen varios tipos de suavizadores en la literatura, algunos más robustos que otros. En este estudio se optó por utilizar el suavizador splines que se basa en la idea de la verosimilitud penalizada. En [11] se da la justificación de la fórmula modificada de mínimos cuadrados que utiliza este suavizador, con la particularidad de que el dominio de las funciones es $[0, 1]$. La versión más común de la fórmula es:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2 + \lambda \int \left[f''\left(x\right)\right]^2 dx \tag{3}$$

Una solución a la función modificada de mínimos cuadrados fue encontrada por Reinsch en 1967 y de Boore en 1987: el spline cúbico [1]; haciendo que lo que se requiera sea estimar los coeficientes de los polinomios.

Para la estimación de coeficientes del modelo, se propone el algoritmo de ajuste posterior (“backfitting”) o de Gauss-Seidel, explicado en [1]. Este algoritmo es un caso particular de los métodos iterativos conocidos como de sobre - relajación sucesiva (SOR por sus siglas en inglés).

La escogencia del λ se puede hacer con validación cruzada generalizada [3].

2.3 Modelos lineales mixtos (LMM)

Son regresiones en las que ciertos parámetros tienen submodelos probabilísticos. Es una fusión de los modelos lineales generalizados y los efectos aleatorios que se explican más adelante. Rosenberg y Guszcza en [6] mencionan que son útiles cuando los datos están agrupados en una o más dimensiones.

El modelo es:

$$\overrightarrow{y} = X \cdot \overrightarrow{\beta} + Z \cdot \overrightarrow{u} + \overrightarrow{\varepsilon}, \tag{4}$$

$\overrightarrow{\beta}$ está vinculado a los efectos fijos, al igual que en los modelos OLS y GLM. El vector $\overrightarrow{u}$ se denomina como de efectos aleatorios, que da la estructura de heterogeneidad inter e intra clase, y se asume que tiene media cero y varianza dada por la matriz D, por lo que se dice que tiene una distribución $\overrightarrow{u} \sim \left(\overrightarrow{0}, D\right)$. Mientras que $\overrightarrow{\varepsilon} \in \mathbb{R}^N$, es el vector de errores, que también tiene media cero, pero la matriz de covarianzas es Σ, por lo que $\overrightarrow{\varepsilon} \sim \left(\overrightarrow{0}, \Sigma\right)$. Otro supuesto es que $\overrightarrow{u}$ es independiente de $\overrightarrow{\varepsilon}$.

Una opción para estimar los parámetros es usar verosimilitud restringida (REML)[1]. Tiene como ventaja que toma en cuenta los grados de libertad usados para la estimación de los efectos fijos.

[1]“Este procedimiento compensa la pérdida de grados de libertad que resulta de la estimación de los efectos fijos y produce estimaciones menos sesgadas de las componentes de varianza” [2].

3 Datos utilizados

Se analizó la cartera de un seguro que ampara a las personas en caso de lesiones por accidentes o tratamientos por enfermedades, entre otras causas similares. Se obtuvo de los sistemas informáticos una consulta de los datos almacenados en los años 2008-2015 de estos asegurados en cuanto a aseguramiento y siniestros (montos pagados a los asegurados).

Las exposiciones de los riesgos son la fracción de año que estuvo el asegurado en la póliza. Para que los siniestros fueran comparables, se trajeron a valor presente con el índice de precios del Banco Central de Costa Rica. Luego, se acotaron por el monto asegurado individual, ya que ningún siniestro lo puede sobrepasar.

Al tratarse de un seguro, es de esperar que no todas las personas presenten reclamos en todos los años - póliza. Es por ello, que se cruzó la base de siniestros con la de cartera expuesta. Quienes no tuvieron siniestros en algún año en particular, se les asignó el valor de cero en la variable dependiente siniestro anual.

Se menciona que en algunas variables cualitativas se renombraron categorías, mientras que a dos predictores continuos y la respuesta se les hizo cambio de escala por motivo de confidencialidad de la información.

Los datos se dividieron en una muestra para ajustar los modelos considerando los años 2008-2012, mientras que, para validar la calidad predictiva, se usó el complemento, (2013-2015).

Para este estudio se estableció:
Variable respuesta o salida y_{ij}: es la suma de los siniestros ocurridos pagados en el año $-$ póliza j a nivel de persona i, netos de la participación del asegurado (coaseguro y límite anual). Es una variable de tipo mixto, con una masa de probabilidad cero, y continua en $\mathbb{R}^+$.
Variables predictivas candidatas: línea, tipo de cédula, sexo, edad, parentesco, núcleo familiar, monto asegurado, recargo de selección, rango de selección, antigüedad, rango de antigüedad, exposición, rango de exposición, año del siniestro, tamaño del grupo, rango del tamaño. Se aclara que el recargo de selección se aplica a las personas que tienen índices de salud por encima de los estándares normales (condición de subnormal).
Las variables que corresponden a rangos de variables cuantitativas se crearon con el propósito de dividir las observaciones en subgrupos de tamaño similar y facilitar el proceso de suscripción.

3.1 Inferencia estadística

La Tabla 1 muestra los estadísticos de las variables cuantitivas, donde se puede ver, por ejemplo, que la mayoría de personas están en edad productiva, y tienen en promedio casi 10 años de tener el seguro. La Tabla 2 muestra la distribución de los individuos en cada variable cualitativa, con la particularidad de que la primer categoría de cada una es la que concentra la mayoría de ellos.

Tabla 1: Estadísticos de las variables cuantitativas de la muestra total.

Estadístico	**Variable**							
	Edad	Antig	Expos_anual	Rec_sel	Monto_aseg	Ann_sin	Sin_anual	Tam_grupo
Mínimo	0	0,08	0,0	0	0	6	0	1
Percentil 25%	20	7,08	0,44	0	5.000	6	0	11
Media	34,07	9,87	0,74	1,18	6.143,07	8,04	212,35	135,2
Percentil 75%	47	13,08	1,0	0	7.500	9	90,90	116
Máximo	91	15,33	1,0	300	10.000	13	10.000	994
Mediana	34	9,33	1,0	0	5.000	8	0	41
Desv. estánd.	18,01	3,72	0,39	7,5	2.271,95	1.86	753,56	235,88

Tabla 2: Distribución de los individuos según categorías, muestra total.

Línea	Tipo de cédula	Sexo	Parentesco	Núcleo
L1: 50.230	C1: 55.946	S1: 29.777	P1: 34.720	N1: 20.869
L2: 14.735	C2: 9.019	S2: 35.188	P2: 11.106	N2: 12.167
			P3: 19.139	N4: 8.085
				N5: 7.226
				N3: 6.920
				N6: 3.770
				Otros: 5.928

En el gráfico de la densidad de los siniestros de la muestra de ajuste (Figura 1), se ve que hay una masa de probabilidad entre 0 y 1.000, indicando una asimetría hacia la izquierda por la presencia de valores extremos y muchas observaciones iguales a cero, tal y como lo describen la curtosis de $60,76$ y el sesgo de $6,7251$.

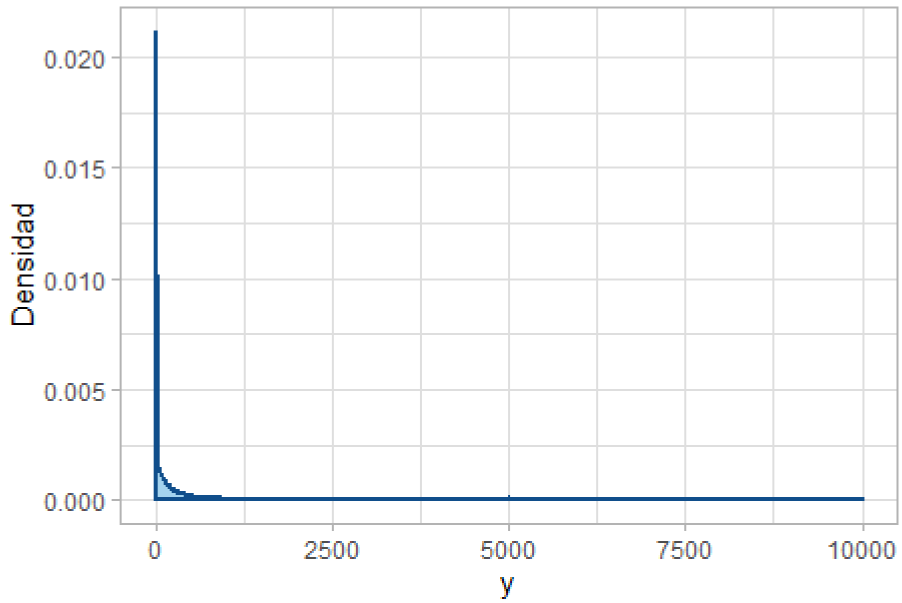
Se deduce que los modelos lineales ordinarios no serían apropiados, puesto que estos suponen una distribución normal de la variable dependiente. Una transformación como es el logaritmo de los siniestros podría solventar el problema, con la implicación de que se deben eliminar las observaciones iguales a cero.

Por último, se menciona que en la base original había 15 tipos de núcleos familiares, pero muchos tenían pocas observaciones, tal y como se vio en la Tabla 2. En la estimación de modelos estos núcleos poco importantes arrojaban valores p muy altos, y al mismo tiempo, tomando en cuenta que en el proceso de suscripción es deseable que existan pocas categorías a escoger en las variables cualitativas, se aplicó el método de k-medias a los siniestros promedio observados en los tres tipos de parentesco, mediante el uso de la función *kmeans* del paquete *stats*. Se determinó que los 15 núcleos se podrían agrupar en 4.

Se trató de hacer una agrupación que fuera balanceada y que los miembros de cada subconjunto tuvieran semejanzas. La descripción de los grupos y el peso que tienen en la muestra se dan a continuación.

- Grupo A. 31%. Personas P1 aseguradas individualmente.
- Grupo B. 45%. Las familias tienen miembros con parentesco P2 y P3.
- Grupo D. 13%. No hay presencia de familiares P2.
- Grupo E. 11%. Hay presencia de P2, pero no de P3.

Figura 1: Densidad de los siniestros.



4 Estimación de los modelos

Se utilizó el software estadístico R tanto para el análisis descriptivo de la sección anterior como para el ajuste de los modelos a los datos. Los paquetes utilizados fueron *base*, *broom*, *car*, *coda*, *cowplot*, *gam*, *GGally*, *ggplot2*, *graphics*, *lattice*, *lmtest*, *MCMCglmm*, *mgcv*, *nlme*, *plot3D*, *stats* y *xtable*. A continuación, se describen los resultados obtenidos en cada modelo.

4.1 Modelo lineal ordinario (OLS)

Se usó la función *lm()* del paquete *stats*, en la que se hace la minimización de los errores al cuadrado. En cuanto a la selección de variables los enfoques tradicionales son "hacia adelante" (comenzar con la intercección e ir incluyendo variables) o "hacia atrás". En este último lo que se hace es eliminar progresivamente variables por valor p superior a un nivel de confianza, como puede ser 5%. Existen otros mecanismos automáticos como la regresión Lasso que buscan maximizar la verosimilitud penalizada. Se hizo primeramente el ejercicio de selección "hacia atrás" y se comparó con el obtenido con Lasso. Como no hubo una mejoría significativa en el periodo de prueba 2013 - 2015, se conservaron los primeros resultados.

Se prefirió utilizar rangos en los predictores puesto que mejoraban los criterios R_a^2 y los de información de Akaike (AIC) y Bayes (BIC o de Schwarz). Estos dos últimos se definen de la siguiente forma:

$$AIC := -2 \cdot l\,(\theta, \overrightarrow{y}) + r \tag{5}$$

$$BIC := -2 \cdot l\,(\theta, \overrightarrow{y}) + r \cdot \ln(n) \tag{6}$$

donde r es el número de parámetros del modelo y n es la cantidad de datos. Como se multiplica la log-verosimilitud por una constante negativa, entre mejor sea el ajuste (mayor verosimilitud), menor será el valor del criterio, por lo que se busca el modelo que minimice estos criterios. La constante r consiste en una penalización ya que siempre se busca la parsimonia.

En [12] se menciona que el criterio de Bayes al involucrar tanto r como $\ln(n)$, la medida se separa del tamaño muestral y hace una penalizacián mayor, buscando el modelo más sencillo aunque no sea el más intuitivo. En [10] se presenta la derivación algebraica de ambos criterios.

El modelo que resultó superior se presenta en la Tabla 5 del Anexo con el detalle del ajuste y los coeficientes de regresión.

Para verificar las hipótesis, se hizo un gráfico cuantil-cuantil de los errores y se notó que siguen muy de cerca una distribución normal. Pero, en el gráfico de la función de auto-covarianza (acf) se vio una alta autocorrelación de orden 1. Esto se comprobó al hacer la prueba Durbin-Watson, ya que el estadístico dio 1.9137 con valor $p < 0.001$.

4.2 Modelo aditivo (AM)

Los criterios de información Akaike y de desviación sirvieron para seleccionar el modelo. Este último consiste en dos veces la diferencia del modelo saturado y el que se está valorando.

Al igual que en el modelo OLS, la transformación logarítmica mejoró considerablemente el ajuste.

Con la función *gam* del paquete *mgcv* se probó aplicar splines a algunas variables, pero solo con la antigüedad se logró una mejora. El modelo ajustado está en la Tabla 6 del Anexo. El método convergió después de 200 iteraciones.

En general, el comportamiento de los errores fue normal como era de esperar. Los gráficos de residuos contra predictores mostraron un comportamiento aleatorio, excepto con el recargo de selección.

4.3 Modelo lineal mixto (LMM)

Se recurrió a la función *lme* en el paquete *nlme* que permite ajustar este tipo de modelo. Se procedió de forma similar que con los modelos anteriores, se valoraron los criterios AIC y BIC, pero también, se tomó en cuenta que la varianza entre categorías del efecto aleatorio fuera más pequeña que la varianza residual, ya que favorece la predicción tal y como lo indica [9].

Como este tipo de modelo permite definir niveles para estimar siniestros, se probaron varias combinaciones, teniendo en cuenta que fueran razonables para la industria.

No existió de antemano una recomendación de cual estructura podría ser adecuada para modelar siniestros. Se probaron varias alternativas de estructuras de correlación, pero se vio que aportaban poco o empeoraban el ajuste. Numéricamente se constató haciendo una prueba ANOVA.

La transformación logarítmica de la respuesta también fue necesaria en este modelo. Los efectos aleatorios que resultaron relevantes fueron linea + monto_aseg + rango_sel.

En la Tabla 7 se pueden ver las estadísticas principales del ajuste. Solo la categoría E de la variable nucl_reclas2 se ve poco significativa, pero como las otras sí aportan al modelo, no se eliminó el predictor. Los coeficientes de los efectos fijos se muestran al final de la Tabla.

5 Comparativo de los modelos

Para comparar la calidad los modelos en el periodo de ajuste se buscaron criterios que fueran comunes, siendo AIC y BIC los que se repetían, los cuales fueron explicado en la sección 4. En la tabla 3 se muestra que el OLS fue ligeramente superior.

Tabla 3: Comparativo de modelos basado en criterios.

Tipo	**Número de modelo**	**AIC**	**BIC**	R^2	**Desviación**
$\ln(y)$ OLS	11	59.209,54	59.397,65	20,30%	
$\ln(y)$ AM Splines	7	58.562,46	58.750,76		13,21
$\ln(y)$ LMM[1]	25	59.355,10	59.472,66		

1 No se puede estimar la desviación cuando se aplica el método REML

Otro ejercicio realizado fue el pronóstico en un periodo posterior al de calibración para valorar el desempeño de los modelos. Se hizo la estimación de los siniestros de los 2.733 individuos que estuvieron expuestos en los años 2013-2015 y se comparó con los datos reales.

El RMSE calculado para cada modelo fue OLS: 1.149.034 y AM: 1.152.882. No se pudo calcular para el LMM porque uno de los pronósticos resultó inválido, puesto que para las variables del efecto aleatorio tenía tipo de línea L2, monto asegurado 7.500 y rango de selección $101 - 110$. Como esta combinación no existe en los datos de ajuste, el vector de coeficientes de efectos aleatorios no tenía un valor a asignar para este pronóstico.

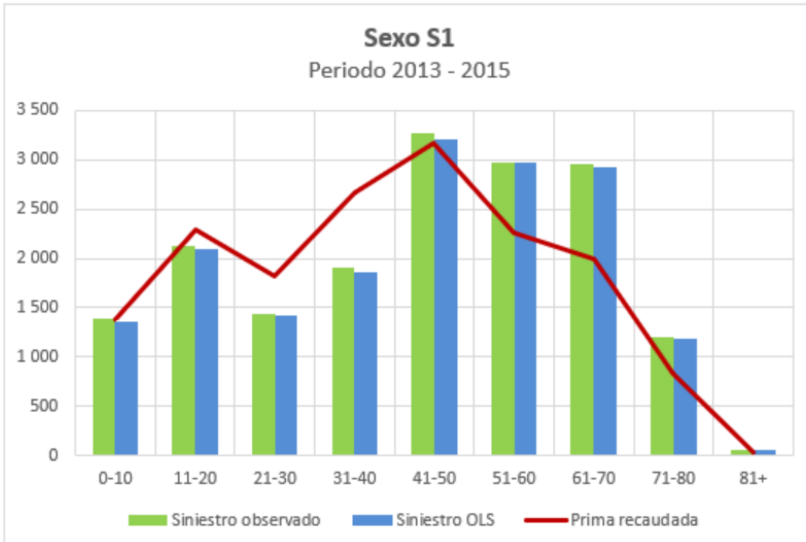
En la Tabla 4 se ve el dato de la suma del logaritmo del siniestro por rango de edad y sexo, tanto el real como el estimado. En los tres se ve que la diferencia porcentual es de alrededor del 1%.

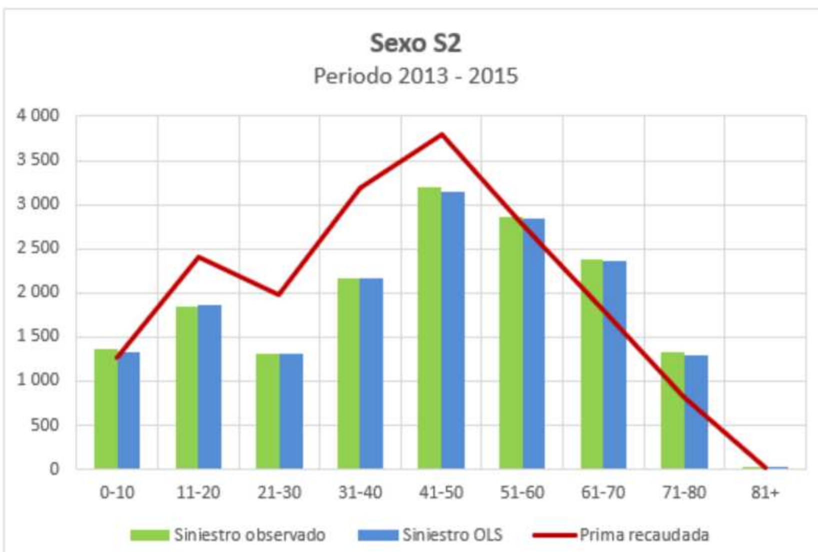
Tabla 4: Comparativo del $\ln(y)$ observado y estimado por rango de edad y sexo.

	Siniestro observado		**Siniestro pron. OLS**		**Siniestro pron. AM**		**Siniestro pron. LMM**	
	Sexo		**Sexo**		**Sexo**		**Sexo**	
Rango de edad	S1	S2	S1	S2	S1	S2	S1	S2
0-10	1 388	1 359	1 356	1 333	1 370	1 344	1 357	1 333
11-20	2 131	1 845	2 093	1 870	2 098	1 873	2 095	1 871
21-30	1 443	1 318	1 417	1 308	1 420	1 308	1 417	1 305
31-40	1 900	2 172	1 867	2 168	1 874	2 172	1 868	2 167
41-50	3 272	3 200	3 207	3 149	3 201	3 144	3 205	3 136
51-60	2 976	2 855	2 969	2 850	2 958	2 838	2 968	2 849
61-70	2 962	2 381	2 928	2 360	2 913	2 347	2 932	2 361
71-80	1 206	1 327	1 184	1 297	1 180	1 292	1 185	1 300
81+	57	28	53	27	53	27	53	27
Total	17 335	16 484	17 074	16 361	17 067	16 346	17 081	16 348
Siniestro total		33 819		33 435		33 413		33 429
% diferencia				-1,1%		-1,2%		-1,2%

Desde otra perspectiva, se hizo la comparación del siniestro total observado por rango de edad y sexo en el periodo de prueba 2013 - 2015. En el gráfico 2 se ve que el modelo OLS aproxima con mucha precisión el monto del siniestro por rango de edad observado en ese periodo, mientras que la prima pura recaudada de los clientes en esos años (línea roja del gráfico) resultó insuficiente en algunos rangos de edad, y en otros, excesiva, implicando que las tarifas empleadas por la aseguradora no fueron las más apropiadas.

Figura 2: Comparación de siniestros estimados con el modelo OLS, siniestros observados y prima pura recaudada para los sexos S1 y S2. Periodo 2013 - 2015





Comparación con otros ejemplos de la industria

Se investigó si habían estudios previos en los que se hubiesen ajustado modelos a seguros de personas (accidentes y gastos médicos). Se encontró que ha sido más habitual que se utilice en seguros de daños como el de automóviles donde la variable de ubicación geográfica tiene mucha relevancia. Sin embargo fue posible encontrar dos que se explican brevemente de seguido.

En el capítulo 2 de [6] Rosenberg y Guszcza describen un modelo lineal para los gastos médicos anuales por persona diagnosticada con diabetes utilizando datos MEPS (Medical Expenditure Panel Survey) de Estados Unidos. Ellos aplicaron transformación logarítmica al dato de siniestro anual al igual que se hizo en este estudio.

Se resalta que la edad tuvo alta importancia en el valor de siniestro. Otro elemento a destacar es que las enfermedades preexistentes como las coronarias, colesterol y otros no son las más influyentes, similar a lo sucedido con el recargo de selección de riesgos. Esto podría deberse a que el seguro se utilice con mucha frecuencia para atender accidentes o padecimientos leves.

Se destaca que obtuvieron un $R^2 = 0,2618$, mientras que en el Tabla 5 de este estudio se ve que fue $R^2 = 0,203$. Se puede decir que a pesar de que se trata de modelos con muchas variables explicativas y volumen de datos, es difícil alcanzar a explicar la varianza en un porcentaje muy alto, es decir, buscar que R^2 sea cercano a 1.

También del libro [6], el investigador Peng Shi presentó un modelo para el costo las consultas médicas de personas que únicamente cuentan con seguro privado usando los datos MEPS del 2008. Se centró en la obtención de los valores positivos, dejando de lado la frecuencia. En esta práctica se debió recurrir a lo mismo porque la transformación logarítmica de y obligó analizar solo valores positivos.

Se nota que la edad y el sexo femenino tuvieron una correlación positiva. Lo mismo se pudo observar en el presente estudio (ver Tabla 5). Situación diferente se presentó con la condición de salud, ya que sí resultó relevante.

6 Discusión y resultados

La investigación bibliográfica permitió llevar a cabo hallazgos relevantes para implementar la técnica de modelación de siniestros, ya que se identificaron variables disponibles que podían ser útiles como predictores. Al mismo tiempo, en el planteamiento de los modelos se tomó en consideración transformaciones y jerarquías de variables vistas en las fuentes consultadas que favorecieron el ajuste de los modelos al tipo de riesgo estudiado.

En el ejercicio realizado, la Figura 2 mostró que el modelo OLS hizo una mejor estimación del siniestro por rango de edad y sexo que el modelo univariado que originó las primas cobradas (línea roja). Los tres modelos estimaron los siniestros en el periodo 2013 - 2015 con una desviación porcentual de alrededor del 1%. Al compararlos se ve que el modelo OLS es mejor en cuanto a los criterios AIC, BIC y RMSE, pero por un margen muy pequeño.

El modelo lineal mixto tiene la cualidad de que se pueden establecer jerarquías. Esto permite tener un mayor control de varianza, ya que se espera que individuos dentro un mismo grupo tenga poca volatilidad, pero que la varianza intergrupal sea grande. La desventaja se presentó cuando se trató de pronosticar un siniestro con una combinación de características que no se habían visto en el periodo de muestra. Con los otros dos sí fue posible, demostrando que un modelo estadístico multivariado sí se puede utilizar en el cálculo de primas para grupos con poca o nula información.

Hay variables que están presentes en los tres modelos: edad, sexo, monto asegurado, parentesco, núcleo familiar, antigüedad y año del siniestro. Este resultado puede ser de utilidad para seleccionar qué información es realmente indispensable obtener de los asegurados para hacer una tarificación adecuada del riesgo y al mismo tiempo buscando que la suscripción sea lo más simplificada posible.

Otro elemento a destacar es que las variables antigüedad y año del siniestro colaboraron el pronóstico, lo que permite decir que se podría establecer un esquema de primas dinámico que actualice las mismas de acuerdo a la antigüedad de la persona en la póliza así como del año calendario.

Como parte de las ideas originales estaba encontrar un modelo que englobara la frecuencia y la severidad, pero se debió encauzar el trabajo a solo ajustar esta última en búsqueda de una buena calidad de las estimaciones, por el hecho de aplicar una transformación de la respuesta. En el futuro se podría incursionar en probar modelos de dos partes para distribuciones con muchos ceros, como es la regresión logística para determinar la probabilidad de ocurrencia de los eventos.

Agradecimientos

Agradezco al Instituto Nacional de Seguros por haber brindado la oportunidad de realizar este trabajo.

Financiamiento

Este trabajo ha sido parcialmente financiado por el Instituto Nacional de Seguros.

Referencias

[1] A. Buja, T. Hastie, R. Tibshirani, *Linear smoothers and additive models*, The Annals of Statistics **17**(1989), no. 2, 453–510. Doi: 10.1214/aos/1176347115.

[2] J. Correa Morales, J. Salazar Uribe, *Introducción a los modelos mixtos*, Universidad Nacional de Colombia, Bogotá, 2016. En: http://bdigital.unal.edu.co/57330/1/introduccionalosmodelosmixtos.2016.pdf. Consultado el 07/09/2018, 10 a.m. Doi:10.1214/aos/1176344136.

[3] M. Durbán, "Métodos de suavizado eficientes con P-splines". Universidad Carlos III, Madrid, España, 2015. En: http://ciencias.bogota.unal.edu.co/fileadmin/content/eventos/simposioestadistica/documentos/memorias/MEMORIAS_2015/Cursillos/Durban_Metodos_Suavizado_P-spline.pdf. Consultado el 11/09/2018, 3:00 p.m.

[4] J.J. Faraway, *Linear models with R*, no.1, Chapman & Hall/CRC, Boca Raton FL, 2004. Doi: https://doi.org/10.4324/9780203507278.

[5] J.J. Faraway, *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression*, no.1, Chapman & Hall/CRC, Boca Raton FL, 2006. Doi: 10.1201/9781315382722.

[6] W.E. Frees, R.A. Derrig, G. Meyers, *Predictive Modeling Applications in Actuarial Science: Predictive Modeling Techniques*, Cambridge University Press, **1**, 2014. Doi: 10.1017/CBO9781139342674.

[7] W.E. Frees, R.A. Derrig, G. Meyers, *Predictive Modeling Applications in Actuarial Science: Case Studies in Insurance*, Cambridge University Press, **2**, 2016. Doi: 10.1017/CBO9781139342681.

[8] P. Ibarrola, "Gauss y la estadística", 2017. En: https://fme.upc.edu/ca/arxius/butlleti-digital/gauss/060215_conferencia_ibarrola.pdf. Consultado el 03/08/2019, 5:00 p.m.

[9] K.H. Jensen, "Linear mixed effects models (lme)", 2016. En: https://folk.uib.no/nzlkj/psychR/day4/04_lme.pdf. Consultado el 22/11/2018, 2:00 p.m.

[10] A. Montesinos López, *Estudio del AIC y BIC en la selección de modelos de vida con datos censurados*, 2011. En: `https://probayestadistica.cimat.mx/sites/default/files/PDFs/TE414MontesinosLopez.pdf` Consultado el 14/10/2020, 7:00 p.m.

[11] D. Nychka, *"Splines as local smoothers", The Annals of Statistics* **23**(1995), no. 4, 1175–1197. Doi: `10.1214/aos/1176324704`.

[12] J.G. Rubalcaba, *Cosas que conviene saber al usar AIC, DIC y otros criterios de información*, 2016. En: `https://jgrubalcaba.wordpress.com/`. Consultado el 14/10/2020, 8:00 p.m.

Anexo. Modelos ajustados

Tabla 5: Modelo lineal ordinario de $\ln(y)$.

Fórmula	lm(formula = log(y) sexo + parent + edad + antig + rango_exp + rango_sel + monto_aseg + an_sin + nucl_reclas2 + rango_tam_grup)						
Coeficientes							
	Estim.	***Error est.***	***Valor t***	***Pr(>|t|)***		***Inf. 95%***	***Sup. 95%***
Intersección	10,58	0,07	151,17	<0,001	***	10,44	10,71
sexoS2	0,05	0,02	2,74	0,006	**	0,01	0,09
parentescoP2	0,05	0,03	1,70	0,088	.	-0,01	0,11
parentescoP3	0,46	0,04	12,83	<0,001	***	0,39	0,54
edad	0,03	0,00	36,63	<0,001	***	0,03	0,03
antig	0,01	0,00	5,01	<0,001	***	0,01	0,02
exp 0,26-0,5	0,35	0,06	5,40	<0,001	***	0,22	0,47
exp 0,51-0,75	0,42	0,06	6,55	<0,001	***	0,30	0,55
exp 0,76-1	0,67	0,05	12,98	<0,001	***	0,57	0,78
sel 101-110	0,75	0,26	2,92	0,004	**	0,25	1,26
sel 121-200	1,36	0,48	2,84	0,005	**	0,42	2,30
sel 201+	-0,38	0,68	-0,57	0,570		-1,72	0,95
monto_aseg	<0,001	-	8,92	<0,001	***	0	0
an_sin	-0,04	0,01	-5,52	<0,001	***	-0,05	-0,02
nucl_reclas2B	-0,20	0,03	-7,49	<0,001	***	-0,25	-0,15
nucl_reclas2D	-0,15	0,03	-4,70	<0,001	***	-0,21	-0,09
nucl_reclas2E	0,06	0,03	1,74	0,082	.	-0,01	0,12
tam_gr6-25	-0,22	0,03	-7,67	<0,001	***	-0,27	-0,16
tam_gr26-49	-0,26	0,03	-9,94	<0,001	***	-0,31	-0,21
tam_gr50-99	-0,32	0,03	-11,31	<0,001	***	-0,38	-0,27
tam_gr100-249	-0,43	0,03	-13,66	<0,001	***	-0,49	-0,37
tam_gr250+	-0,31	0,03	-10,48	<0,001	***	-0,37	-0,25
—							
Significancia:	0 "***"	0,001 "**"	0,01 "*"	0,05 "."	0,1 " "	1	
Estadísticos							
Error residual est.	1,17		Grados lib.	18,70			
R^2 múltiple	0,20		R^2 ajust.	0,20			
Estadístico F	228,10		Grados lib.	21 y 18,70			
Valor p	<0,001		log-ver.	-29.582			
AIC	59.210		BIC	59.390			
			Desv.	25.832			

Tabla 6: GAM Splines de $\ln(y)$.

Fórmula	log(y) = sexo + parent + edad + s(antig) + exp_an + rec_sel + (mont_aseg)+ (an_sin) + nucl_recl2 + rang_tam_grup				
Familia:	Gauss. Inv.				
Func. enlace:	$1/\mu^2$				
Coeficientes					
	Estim.	***Error est.***	***Valor*** t	***Pr(>|t|)***	
Intersección	0,00800	<0,001	113,95	<0,001	***
sexoS2	<-0,001	<0,001	-2,67	0,008	**
parentescoP2	<-0,001	<0,001	-1,88	0,06	.
parentescoP3	<-0,001	<0,001	-10,57	<0,001	***
edad	<-0,001	<0,001	-34,91	<0,001	***
exp_anual	-0,00100	<0,001	-14,60	<0,001	***
rec_sel	<-0,001	<0,001	-3,54	<0,001	***
monto_aseg	<-0,001	<0,001	-8,18	<0,001	***
an_sin	0,00004	<0,001	5,58	<0,001	***
nucl_reclas2B	0,00020	<0,001	6,95	<0,001	***
nucl_reclas2D	0,00013	<0,001	3,91	<0,001	***
nucl_reclas2E	-0,00004	<0,001	-1,30	0,20	
tam_gr6-25	0,00020	<0,001	6,98	<0,001	***
tam_gr26-49	0,00030	<0,001	9,36	<0,001	***
tam_gr50-99	0,00032	<0,001	10,89	<0,001	***
tam_gr100-249	0,00044	<0,001	13,22	<0,001	***
tam_gr250+	0,00030	<0,001	10,10	<0,001	***
—					
Significancia:	0 "***"	0.001 "**"	0.01 "*"	0.05 "."	0.1 " "
Significancia aproxim. de términos de suavizam.					
	edf	***Ref.df***	***F***	***Valor*** p	
s(antig)	6,03	7,16	7,62	<0,001	***
Estadísticos					
R^2 ajustado	0,21				
Parám. escala	0,0007	Desv. expl.		21%	
log-verosim	-29.257	Desviac.		13,2	
BIC	58.751				
GVC	0,0007				

Tabla 7: LMM de $\ln(y)$ con efectos aleatorios linea/monto_aseg/rango_sel.

Fórmula	nlme::lme(log(y) = sexo + parent + edad + antig + exp_anual + nucl_reclas2 + an_sin random = linea/ monto aseg/ rango sel)					
Método	REML					
Efectos fijos						
	Estim.	***Error est.***	***Valor t***	***Pr(>|t|)***	***Inf. 95%***	***Sup. 95%***
(Intercept)	10,58	0,145	72,81	-	10,30	10,87
sexoS2	0,06	0,018	3,05	0,002	0,02	0,09
parentescoP2	0,05	0,028	1,90	0,057	<-0,001	0,11
parentescoP3	0,49	0,036	13,65	-	0,42	0,56
edad	0,03	0,001	37,88	-	0,03	0,03
antig	0,01	0,003	4,19	-	0,01	0,02
exp_anual	0,72	0,047	15,43	-	0,63	0,82
nucl_reclas2B	-0,23	0,027	-8,51	-	-0,28	-0,17
nucl_reclas2D	-0,15	0,032	-4,79	-	-0,22	-0,09
nucl_reclas2E	0,05	0,033	1,40	0,160	-0,02	0,11
an_sin	-0,04	0,008	-5,27	-	-0,05	-0,02
Efectos aleatorios						
Fórmula:	linea	monto aseg -linea			rang sel -monto aseg -linea	
	(Inters.)	(Inters.)			(Inters.)	Residuo
Desv. Estánd.	0,18	0,14			0,001	1,18
Residuos intragrupos estandarizados:						
Min	Q1	Med	Q3	Max		
-4,16	-0,69	-0,06	0,59	3,69		
Número de grupos:						
	linea		monto_aseg-linea		rang sel monto aseg -linea	
	2		10			19
Estadísticos						
log-verosim	-29.663		Desviación		n/a	
AIC	59.355		BIC		59.473	