**UNIVERSIDAD SANTO TOMAS**

**FACULTAD DE ESTADÍSTICA**

**APLICACIÓN DE MODELOS LINEALES GENERALIZADOS (GLM) PARA LA TARIFACIÓN TÉCNICA DE SEGUROS DE AUTOMÓVILES EN PREVISORA SEGUROS**

**María Camila Infante González**

**Bogotá, Colombia**

**Agosto 2025**

**APLICACIÓN DE MODELOS LINEALES GENERALIZADOS (GLM) PARA LA TARIFACIÓN TÉCNICA DE SEGUROS DE AUTOMÓVILES EN PREVISORA SEGUROS**

**Planteamiento del problema**

En el mercado asegurador colombiano, la línea de seguros de automóviles representa una de las más relevantes tanto en volumen de primas como en siniestralidad. Sin embargo, la fijación de tarifas enfrenta diversos desafíos: alta competencia entre aseguradoras, cambios en la frecuencia y severidad de los siniestros, la heterogeneidad del parque automotor, así como la influencia de factores socioeconómicos, geográficos y de comportamiento de los asegurados. En este contexto, contar con metodologías de tarifación que sean robustas, objetivas y adaptables resulta fundamental para garantizar la sostenibilidad técnica y financiera de las pólizas.

Tradicionalmente, los esquemas tarifarios se han basado en tablas de factores o ajustes agregados que, aunque prácticos, pueden llevar a subestimaciones o sobreestimaciones del riesgo. Esto genera distorsiones en la equidad tarifaria, pues asegurados con perfiles de riesgo distintos pueden terminar pagando primas similares. Además, esta aproximación limita la capacidad de la aseguradora para responder con agilidad a la dinámica del mercado y a la evolución del riesgo.

Frente a estas limitaciones, surge la necesidad de adoptar herramientas estadísticas más avanzadas que permitan modelar de manera adecuada la relación entre la frecuencia y severidad de los siniestros con múltiples variables explicativas. Los Modelos Lineales Generalizados (GLM) se han consolidado como el estándar internacional en tarifación actuarial, ya que permiten capturar relaciones no lineales, manejar distribuciones de probabilidad acordes a la naturaleza de los datos (como Poisson o binomial negativa para frecuencia, y Gamma o Inversa Gaussiana para severidad), e incorporar variables categóricas y continuas de manera flexible.

En el caso de Previsora Seguros, aplicar un modelo de este tipo no solo representa una oportunidad para mejorar la precisión en la estimación de primas, sino también para identificar las variables más influyentes en el riesgo de los asegurados, lo cual puede servir de soporte a la definición de políticas tarifarias más técnicas, competitivas y equitativas. De esta manera, el uso de GLM contribuiría directamente a fortalecer la gestión actuarial, la sostenibilidad financiera y la competitividad de la compañía en el ramo de autos.

**Justificación**

La tarifación de seguros de automóviles es un proceso crítico para las aseguradoras, pues de su adecuada estructuración depende tanto la competitividad en el mercado como la sostenibilidad técnica del portafolio. En la actualidad, el uso de metodologías tradicionales basadas en promedios o factores agregados puede conducir a ineficiencias en la asignación del riesgo, lo que repercute en distorsiones tarifarias y posibles desequilibrios financieros.

En este sentido, los Modelos Lineales Generalizados (GLM) ofrecen una alternativa metodológica robusta y ampliamente aceptada en la práctica actuarial internacional, al permitir ajustar la prima de acuerdo con características individuales de los asegurados y de los vehículos, mejorando así la equidad y la exactitud en la estimación de los riesgos.

Para Previsora Seguros, la implementación de un modelo de este tipo se convierte en una oportunidad estratégica: por un lado, permite fortalecer la toma de decisiones técnicas en el diseño tarifario, y por otro, respalda una política de precios más justa y competitiva, acorde con la realidad del mercado y con las mejores prácticas actuariales. Desde el ámbito académico, este proyecto constituye una aplicación directa de los conocimientos adquiridos en la formación estadística, generando productos de valor tanto documentales (informe metodológico y resultados) como instrumentales (modelos y scripts replicables).

En consecuencia, este trabajo no solo aporta al desarrollo profesional del estudiante, sino también al avance de la gestión actuarial de la empresa, evidenciando la importancia de aplicar herramientas estadísticas modernas en problemas reales de la industria aseguradora.

**Objetivo general**

Aplicar Modelos Lineales Generalizados (GLM) para la tarifación del seguro de automóviles en Previsora Seguros, con el fin de mejorar la precisión en la estimación de primas, identificar variables relevantes y apoyar el diseño de una política tarifaria más técnica y equitativa.

**Objetivos específicos**

1. Explorar y depurar la base de datos histórica de siniestros y pólizas de automóviles de Previsora Seguros, garantizando su calidad y consistencia para el modelamiento.
2. Seleccionar y justificar las distribuciones de probabilidad y funciones de enlace adecuadas para modelar la frecuencia y severidad de los siniestros.
3. Construir y ajustar modelos GLM que permitan explicar la variabilidad de la frecuencia y la severidad en función de características del vehículo, del asegurado y de factores externos.
4. Evaluar el desempeño de los modelos propuestos mediante métricas estadísticas y actuariales (bondad de ajuste, devianza, criterios de información y validación cruzada).
5. Proponer una estructura tarifaria basada en los resultados del modelo, que permita mayor equidad y precisión en la estimación de primas.
6. Generar un producto documental y/o instrumental (informe metodológico, scripts en R/Python o tablas tarifarias) que pueda ser utilizado por la Gerencia de Actuaría como insumo en sus procesos de gestión y fijación de precios.

**Marco teórico**

## 1. Situación actual de la tarifación de autos y necesidades del negocio

En el mercado colombiano de seguros de automóviles coexisten (i) alta competencia en precio, (ii) cambios en frecuencia y severidad asociados a factores macroeconómicos, de movilidad y delictivos, y (iii) una marcada heterogeneidad del parque automotor y de los perfiles de conducción. En este contexto, esquemas tradicionales basados en promedios y tablas agregadas pueden generar **inequidad tarifaria** (asegurados con riesgo diferente pagando primas similares) y **riesgo técnico** (prima insuficiente para ciertos segmentos).

De aquí surgen necesidades concretas para Previsora Seguros: - **Mejorar precisión y equidad**: primas alineadas con el riesgo esperado a nivel individual (vehículo–asegurado–uso–territorio). - **Transparencia e interpretabilidad** para decisiones tarifarias y diálogo con negocio y regulador. - **Agilidad** para recalibrar frente a cambios de tendencia, inflación de siniestros, y mezcla de portafolio. - **Gobierno de modelos** que asegure trazabilidad, validación, monitoreo y uso responsable de variables.

Los **Modelos Lineales Generalizados (GLM)** satisfacen estas necesidades al ofrecer un marco estadístico flexible, transparente y probado internacionalmente en tarifación.

## 2. Fundamentos actuariales de la tarifación técnica

### 2.1. Conceptos básicos

* **Exposición (E)**: tiempo o riesgo asegurado (por ejemplo, vehículo–año).
* **Frecuencia (N/E)**: número esperado de siniestros por unidad de exposición.
* **Severidad (S/N)**: costo promedio por siniestro.
* **Costo puro (S/E)**: producto de frecuencia × severidad.
* **Prima técnica**: costo puro ajustado por gastos, margen de seguridad y otros recargos/descuentos comerciales.

La práctica actuarial construye tarifas **multiplicativas**: una *prima base* modulada por **relatividades** (factores) asociados a características del riesgo. Esta estructura es naturalmente compatible con GLM bajo enlace log (aditividad en la escala del predictor lineal ↔ multiplicatividad en la prima).

### 2.2. Segmentación y rating factors

Variables típicas en autos incluyen características del **vehículo** (marca, línea/modelo, clase, cilindraje, año, valor asegurado, potencia o tamaño), del **asegurado/conductor** (edad, antigüedad de licencia, siniestralidad histórica, uso del vehículo), **geografía/territorio** (zona de circulación/garaje), **exposición** (coberturas y deducibles), **canal** y **método de pago**. El objetivo es estimar relatividades estables, monotónicas cuando aplica, y sustentadas en evidencia

## 3. Modelos Lineales Generalizados (GLM): conceptos clave

Un GLM modela ([YX] = = g^{-1}(X)), donde **g(·)** es la función de enlace y **Y** pertenece a la **familia exponencial**. Aspectos esenciales para tarifación:

* **Familias de probabilidad** acordes con la naturaleza de la variable respuesta.
* **Enlace log** para garantizar positividad y facilitar interpretabilidad como relatividades multiplicativas.
* **Offset por exposición**: típicamente ((E)) para modelar tasas (frecuencia por exposición o costo puro por exposición).
* **Estimación por máxima verosimilitud**, inferencia mediante estadísticos de Wald y pruebas de razón de verosimilitud.
* **Medidas de ajuste**: devianza, AIC/BIC, pseudo-(R^2), residuos (Pearson, devianza) y diagnósticos de influencia.

## 4. GLM para frecuencia, severidad y costo puro

### 4.1. Frecuencia de siniestros

* **Poisson (log)**: punto de partida para conteos; asume varianza = media.
* **Quasi-Poisson / Binomial Negativa**: maneja **sobredispersión** (varianza > media) frecuente en siniestros.
* **Modelos con exceso de ceros** (ZIP/ZINB) o *hurdle*: útiles si hay muchos asegurados sin siniestros; se consideran cuando el GLM base no captura bien la masa en cero.

### 4.2. Severidad de siniestros (condicional en haber siniestro)

* **Gamma (log)** o **Inversa Gaussiana (log)** para costos positivos y asimétricos.
* Alternativamente, **Lognormal** (vía GLM aproximado o modelos de dos etapas) si la multiplicatividad log-normal es mejor.
* Tratamiento de **outliers** y **topes/deducibles**: winsorización, modelado de *limited expected value* o componentes por cobertura.

### 4.3. Costo puro / prima pura

* **Tweedie (log, 1<p<2)**: modela simultáneamente masa en cero (sin siniestro) y distribución continua positiva (con siniestro). Es muy usado para prima pura.
* **Modelos en dos partes** (frecuencia × severidad): favorecen interpretabilidad y permiten controles específicos por fenómeno.

### 4.4. Interacciones, no linealidades y parquedad

* Interacciones relevantes (p. ej., **edad del conductor × uso**, **valor del vehículo × marca/clase**, **territorio × uso**).
* Transformaciones y *splines* (p. ej., para edad del vehículo, valor asegurado o antigüedad de licencia) manteniendo interpretabilidad.
* **Parquedad**: preferir especificaciones estables y simples que generalicen bien.

## 5. Ingeniería de variables y tratamiento de datos

* **Calidad de datos**: deduplicación, conciliación de pólizas–siniestros, tratamiento de fechas, coherencia de coberturas y sumas aseguradas.
* **Exposición**: cálculo preciso (pro-rata), manejo de cancelaciones/anulaciones y fraccionamientos.
* **Codificación de categóricas**: *one-hot* con referencia; consolidación de categorías raras; criterios de fusión por similitud del riesgo.
* **Monotonicidad** cuando es razonable (p. ej., mayor potencia → mayor riesgo) para mejorar estabilidad comercial.
* **Evitar *data leakage***: usar solo información disponible en el momento de cotizar.
* **Ajustes por tendencia** (inflación de costos, efectos de calendario) y **estacionalidad** si aplica.

## 6. Estimación, selección de variables y regularización

* **Selección guiada por criterio de información** (AIC/BIC), pruebas de razón de verosimilitud y significancia práctica (impacto en relatividades).
* **Regularización** (LASSO/Ridge/Elastic Net) como apoyo para controlar colinealidad y estabilizar estimaciones cuando hay muchas variables o granularidad elevada.
* **Releveling y centrado**: elegir categorías de referencia con suficiente exposición y significado de negocio.

## 7. Validación, diagnóstico y estabilidad

* **Bondad de ajuste**: devianza, residuos, p-valores globales.
* **Validación predictiva** *out-of-time/out-of-sample*: particiones temporales y *k*-fold.
* **Calibración**: gráficos de observado vs. esperado por deciles de riesgo, por factor y por territorio.
* **Sensibilidad e influencia**: *Cook’s distance*, leverage, análisis de outliers.
* **Estabilidad temporal**: *population stability index* (PSI), *drift* de variables y **shift de mezcla** (portafolio).
* **Backtesting técnico**: razón de siniestralidad (LR) esperada vs. observada por segmento; *hit ratio* y anti-selección tras cambios tarifarios.

## 8. Construcción de la tarifa comercial

A partir de las relatividades GLM: 1. **Definir prima base** (nivelación) y **escala multiplicativa** por factor. 2. **Suavizar y acotar** relatividades (capping/flooring) para asegurar aplicabilidad comercial y evitar saltos bruscos. 3. **Compatibilizar** con coberturas, deducibles y productos (paquetes vs. a la carta). 4. **Ajustar por gastos, comisiones, impuestos y margen** para obtener la **prima comercial**. 5. **Pruebas de impacto**: simulación *what-if* y *pricing impact* sobre el portafolio (ganadores/perdedores, elasticidades). 6. **Política de implementación**: gradualidad, ventanas de monitoreo, y controles de anti-selección.

## 9. Consideraciones de equidad, ética y regulatorio

* **Equidad tarifaria**: evitar el uso de variables sensibles o *proxies* discriminatorios y justificar estadística y éticamente cada factor.
* **Explainability**: GLM facilita interpretar efectos marginales y comunicar criterios a negocio, clientes y supervisor.
* **Gobierno y trazabilidad**: documentación metodológica, versionamiento de datos/modelos/códigos, y bitácora de cambios tarifarios.

## 10. Gobierno del modelo y ciclo de vida

* **Documentación**: supuestos, elección de familia/enlace, variables incluidas, diagnósticos y decisiones de negocio.
* **Reproducibilidad**: *pipelines* en R/Python, ambientes controlados y pruebas automatizadas.
* **Monitoreo continuo**: desviaciones de LR, frecuencia y severidad por segmento, *alertas* de drift, y revisión periódica (por ejemplo, trimestral/semestral).
* **Retraining/recalibración**: criterios y umbrales; *champion–challenger*.

## 11. Limitaciones y riesgos

* **Supuestos de GLM** (correcta especificación de familia/enlace, independencia condicional) pueden vulnerarse en presencia de correlaciones espaciotemporales o **exposición heterogénea**.
* **Colinealidad** y **datos escasos** por celdas finas → relatividades inestables.
* **Cambios estructurales** (normativos, tecnológicos o de movilidad) que alteren patrones de siniestralidad.
* **Sesgos de medición** (subdeclaración, errores en registro de siniestros/valores) y **outliers de severidad**.

## 12. Productos aplicables al cierre del proyecto

* **Informe técnico**: metodología, resultados, diagnósticos y guía de uso de la tarifa.
* **Scripts replicables (R/Python)**: limpieza, modelado GLM, validación y generación de relatividades.
* **Tablas tarifarias/scorecards**: prima base y factores por variable; diccionario de variables y reglas de aplicación.
* **Panel de monitoreo**: indicadores clave (LR esperado/observado, drift, calibración por segmento) y alertas.