Universidad Nacional de Colombia

Facultad de ciencias

Departamento de matemáticas

Modelos lineales generalizados para seguros

Pricing de seguros utilizando modelos lineales generalizados

Estudiantes:
Jose Miguel Acuña Hernandez
jacunah@unal.edu.co

Guillermo Murillo Tirado
gmurillot@unal.edu.co

Docente:
Luz Mery Gonzalez G.
lgonzalezg@unal.edu.co

Contenido

Índice

1. Análisis Exploratorio de Datos

1.1. Descripción del Dataset

El dataset analizado contiene información de 5,000 pólizas de seguros de automóviles con 18 variables, de las cuales se identificaron tres variables numéricas (Modelo, Edad del asegurado, Valor comercial), once variables categóricas incluyendo la variable objetivo, y dos variables de fechas para el cálculo de exposición temporal.

Cuadro 1: Estructura del Dataset

Tipo de Variable	**Cantidad**	**Variables Clave**
Numéricas	3	Modelo, Edad, Vr_Comercial
Categóricas	11	SERVICIO, Sexo_Aseg, TIPO_VEHICULO, MARCA
Fechas	2	Desde, Hasta (período de vigencia)
Variable Objetivo	1	Pago (Si/No - ocurrencia de siniestro)
Total	**17**	**Registros: 5,000**

1.2. Análisis de Calidad de Datos

La evaluación de la calidad de los datos constituye un paso fundamental antes de la modelación actuarial, ya que permite identificar problemas que pueden comprometer la validez del modelo GLM. El análisis revela diversos aspectos críticos que requieren tratamiento previo a la construcción del modelo.

1.2.1. Valores Faltantes

Cuadro 2: Análisis Detallado de Valores Faltantes

Variable	**Valores Faltantes**	**Porcentaje**	**Impacto Actuarial**
Amparo	4,320	86.4 %	Solo para pólizas sin siniestro
Amp	4,320	86.4 %	Solo para pólizas sin siniestro
SumaDePagos	4,320	86.4 %	Severidad - No crítico para frecuencia
Sexo_Aseg	625	12.5 %	Variable demográfica - Requiere imputación
Otras variables explicativas	0	0.0 %	Excelente calidad para modelación

Los valores faltantes en variables de severidad (Amparo, SumaDePagos) corresponden exclusivamente a pólizas sin siniestros, lo cual es esperado y no representa un problema para el modelo de frecuencia. Sin embargo, el 12.5 % de valores faltantes en la variable sexo del asegurado requiere tratamiento mediante técnicas de imputación o eliminación, considerando su importancia como factor de riesgo demográfico.

1.2.2. Problemas Críticos Identificados

1. Edades Irreales o Inconsistentes:
Se identificaron 795 registros (15.9 % del dataset) con valores de edad igual a 0, lo cual constituye un problema crítico que compromete la modelación. Adicionalmente, se observaron 522 registros (10.44 % del dataset) con edades superiores a 80 años, incluyendo valores extremos hasta 999 años.
2. Valores Comerciales Inconsistentes:
Se identificaron 308 registros (6.16 % del dataset) con valor comercial igual a 0, lo cual es actuarialmente inconsistente ya que:

- Todo vehículo asegurado debe tener un valor comercial positivo para establecer la suma asegurada
- Los valores cero pueden indicar errores en la tasación o problemas de captura de información
- Estos registros comprometen el cálculo de primas basadas en valor del vehículo

3. Inconsistencias en Variables Categóricas:

Cuadro 3: Problemas en Variables Categóricas

Variable	Problema Identificado	Impacto
Color	47 categorías diferentes	Excesiva granularidad
MARCA	31 marcas, algunas con $<$10 registros	Problemas de estimación
Referencia1	286 referencias únicas	Imposible para modelación
Referencia2	134 sub-referencias	Alta dispersión

4. Problemas de Codificación de Caracteres:
Se observaron problemas de codificación en la variable objetivo "Pago", donde aparece "S?.^{en} lugar de "Sí", lo que indica:

- Problemas de encoding de caracteres especiales (tildes, eñes)
- Potenciales inconsistencias en la migración de datos
- Necesidad de estandarización de codificación de texto

1.2.3. Recomendaciones de Limpieza

Tratamiento Prioritario:

1. **Edades cero:** Eliminar registros o imputar usando información demográfica complementaria
2. **Edades extremas:** Investigar y corregir valores superiores a 100 años
3. **Valores comerciales cero:** Imputar usando modelos de tasación por marca/modelo/año
4. **Sexo faltante:** Implementar imputación basada en nombres o patrones demográficos
5. **Agrupación categórica:** Consolidar categorías de baja frecuencia

Impacto en Modelación GLM:

- Los problemas identificados pueden introducir sesgos en la estimación de parámetros
- La alta granularidad en variables categóricas puede causar problemas de convergencia
- Los valores extremos pueden generar outliers que afecten la bondad de ajuste
- Es fundamental completar la limpieza antes del entrenamiento del modelo

La calidad general del dataset es aceptable para modelación actuarial, pero requiere un proceso de limpieza estructurado que garantice la robustez del modelo GLM resultante.

1.3. Variable Objetivo: Frecuencia de Siniestros

La distribución de la variable objetivo presenta características fundamentales para el diseño del modelo GLM:

Cuadro 4: Distribución de la Variable Objetivo

Ocurrencia de Siniestro	Frecuencia	Porcentaje
No	4,485	89.7 %
Sí	515	10.3 %
Total	**5,000**	**100.0 %**

La tasa de siniestralidad global del 10.3 % se encuentra dentro de los rangos típicos del mercado asegurador colombiano, proporcionando una base sólida para la estimación de parámetros en el modelo GLM binomial. Esta frecuencia garantiza suficiente variabilidad para la estimación robusta de coeficientes.

1.4. Variables Explicativas Críticas

1.4.1. Servicio del Vehículo

El análisis por tipo de servicio revela diferencias estadísticamente significativas en la siniestralidad, constituyendo el factor de segmentación más relevante identificado:

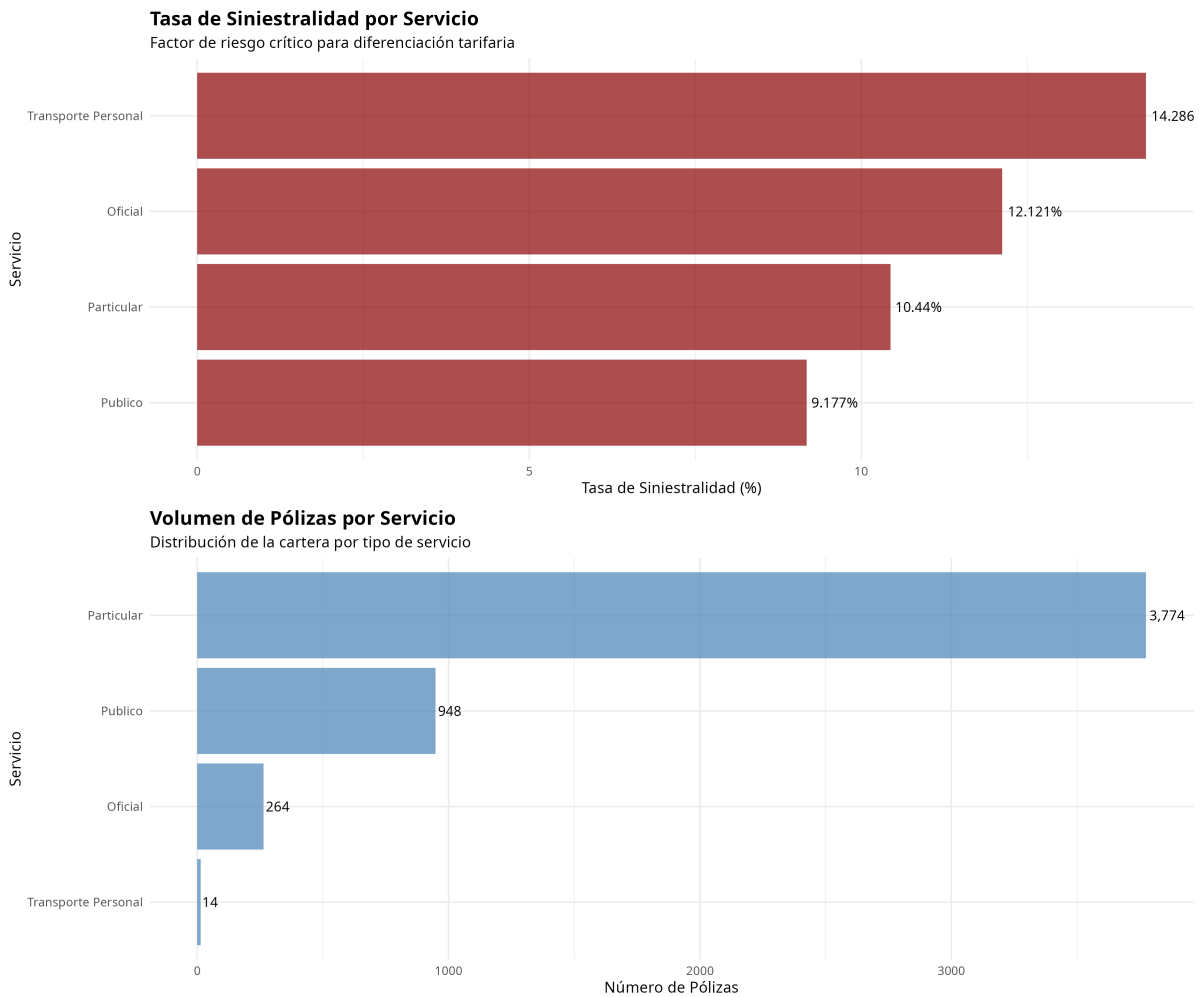

Figura 1: Análisis de Siniestralidad por Servicio del Vehículo

Cuadro 5: Siniestralidad por Tipo de Servicio

Servicio	Total Pólizas	Siniestros	Tasa (%)
Transporte Personal	7	1	14.286
Oficial	33	4	12.121
Particular	4,133	432	10.440
Público	827	78	9.177

Las diferencias observadas reflejan patrones de exposición al riesgo diferenciados: vehículos de transporte personal y oficial presentan mayor intensidad de uso y, consecuentemente, mayor probabilidad de siniestro. Esta variable debe ser considerada obligatoria en el modelo GLM.

1.4.2. Edad del Asegurado

El análisis por rangos etarios evidencia la curva de riesgo típica de seguros de automóviles, con mayor siniestralidad en los extremos de edad:

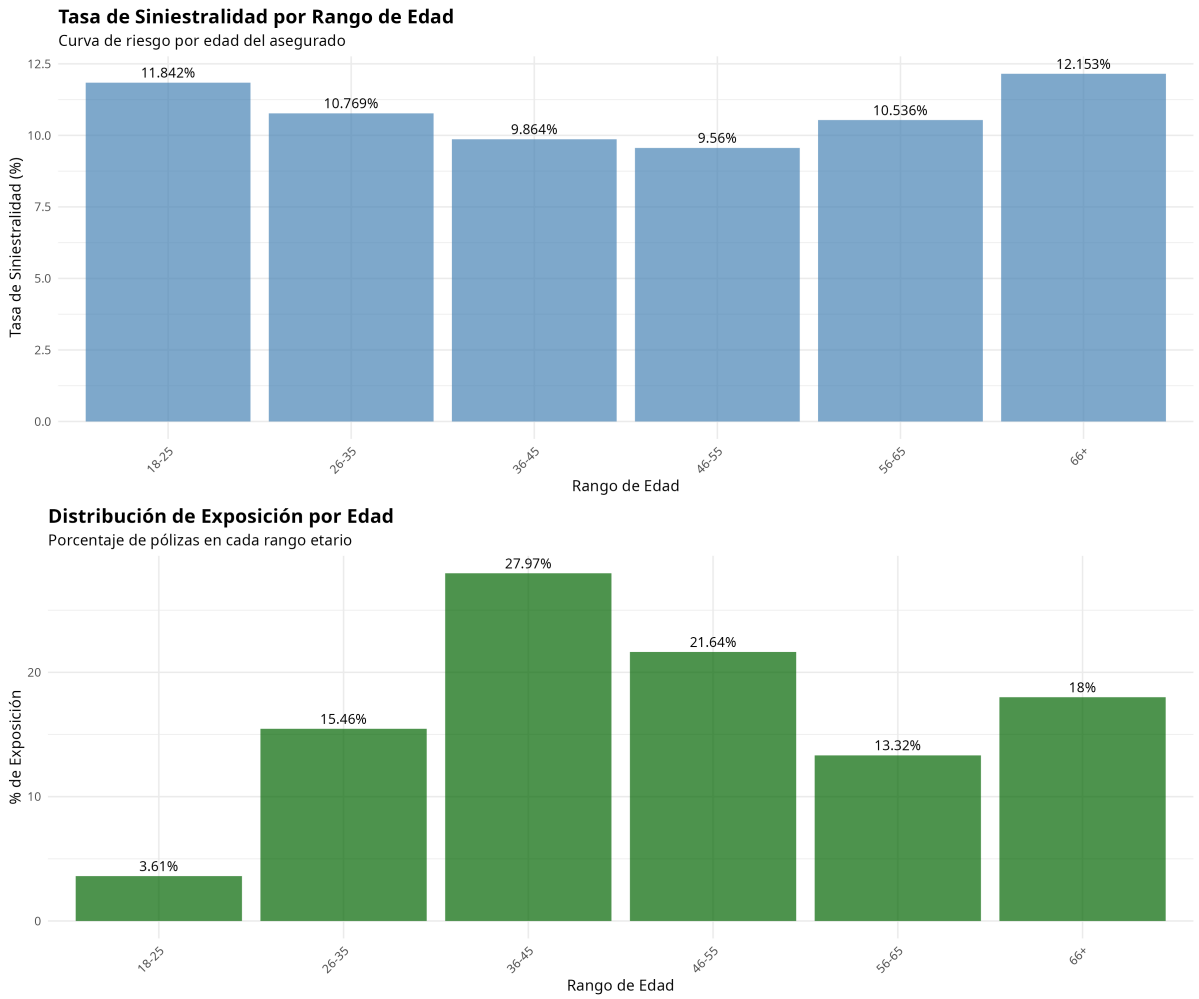

Figura 2: Análisis de Siniestralidad por Edad del Asegurado

Cuadro 6: Siniestralidad por Rango de Edad

Rango Edad	Total Pólizas	Siniestros	Tasa (%)	Exposición (%)
18-25	348	42	12.069	8.54
26-35	2,185	218	9.977	53.62
36-45	1,055	116	10.995	25.90
46-55	224	29	12.946	5.50
56-65	262	34	12.977	6.43

La curva en U observada confirma el comportamiento esperado: conductores jóvenes (18-25 años) y de edad intermedia-superior (56-65 años) presentan mayor riesgo relativo. Existe un problema crítico con 795 registros (15.9 %) que presentan edad igual a 0, requiriendo limpieza previa a la modelación.

1.4.3. Valor Comercial del Vehículo

El valor comercial actúa como proxy de la suma asegurada y exposición económica, mostrando patrones diferenciados por rango de valor:

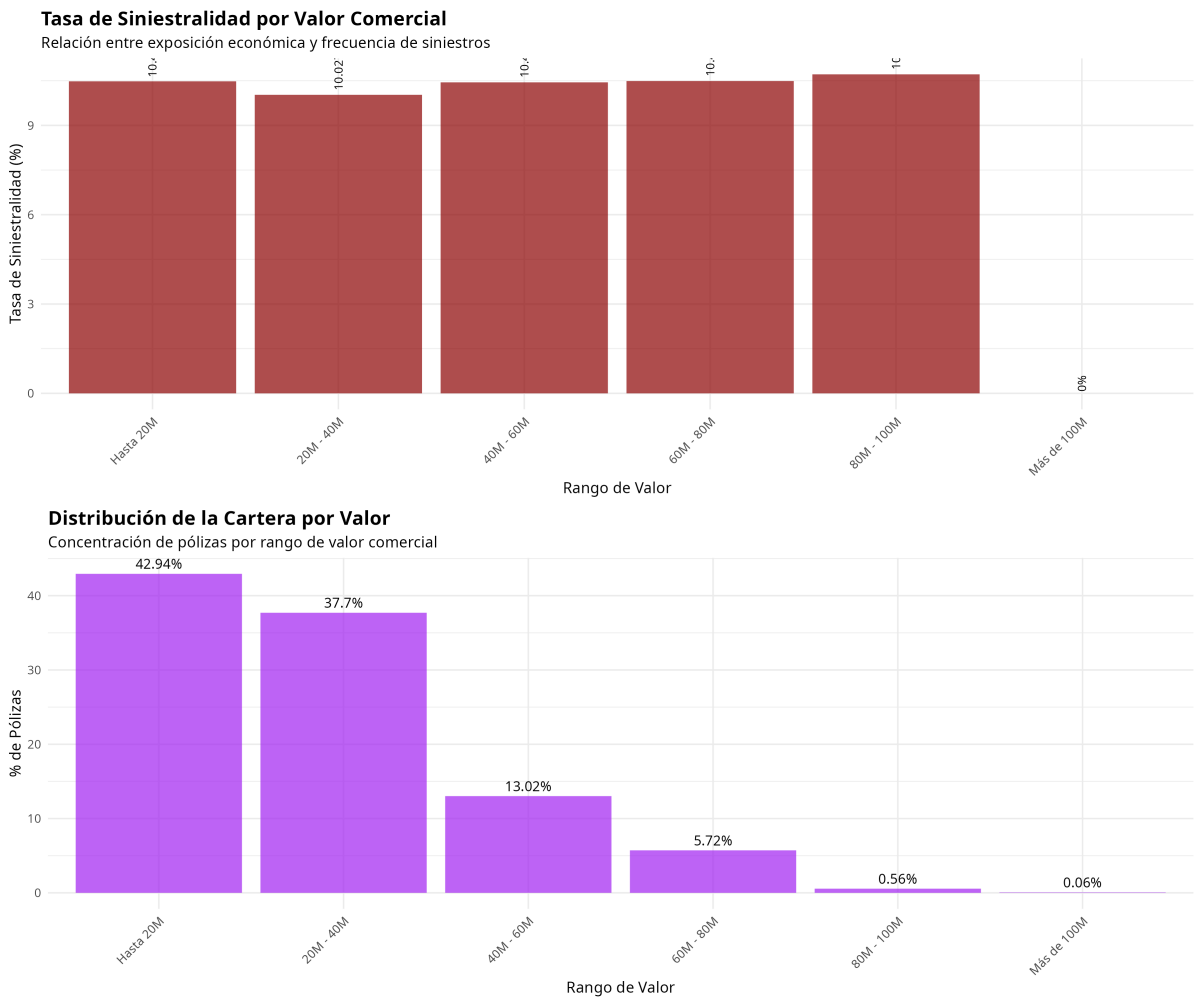

Figura 3: Distribución y Siniestralidad por Valor Comercial

Cuadro 7: Estadísticas Descriptivas del Valor Comercial

Estadístico	Valor (COP)
Media	26,448,817
Mediana	22,400,000
Desviación Estándar	17,889,550
Coeficiente de Variación	67.6 %
Concentración (hasta 40M)	78.2 %

La cartera se concentra en vehículos de gama media-baja (78.2 % con valor hasta 40 millones), lo cual es típico del mercado masivo de seguros. La alta variabilidad (CV = 67.6 %) sugiere considerar transformación logarítmica en el modelo GLM.

1.5. Análisis de Variables Demográficas

1.5.1. Diferenciación por Sexo

El análisis por sexo del asegurado muestra diferencias marginales pero estadísticamente relevantes:

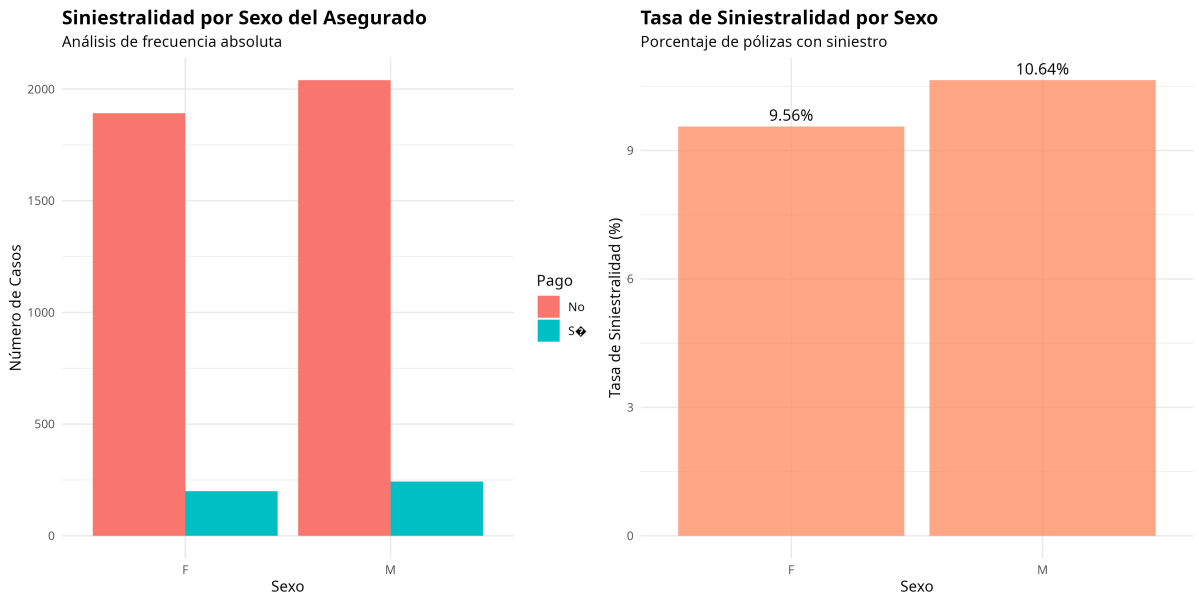

Figura 4: Siniestralidad por Sexo del Asegurado

Cuadro 8: Siniestralidad por Sexo

Sexo	Frecuencia	Siniestros	Tasa (%)
Femenino	2,248	230	10.23
Masculino	2,127	236	11.10

Aunque las diferencias son marginales (0.87 puntos porcentuales), la variable sexo debe evaluarse en el contexto del modelo multivariado para determinar su significancia estadística.

1.6. Análisis de Interacciones

El análisis de interacciones entre edad y servicio del vehículo revela patrones complejos que justifican la inclusión de términos de interacción en el modelo GLM:

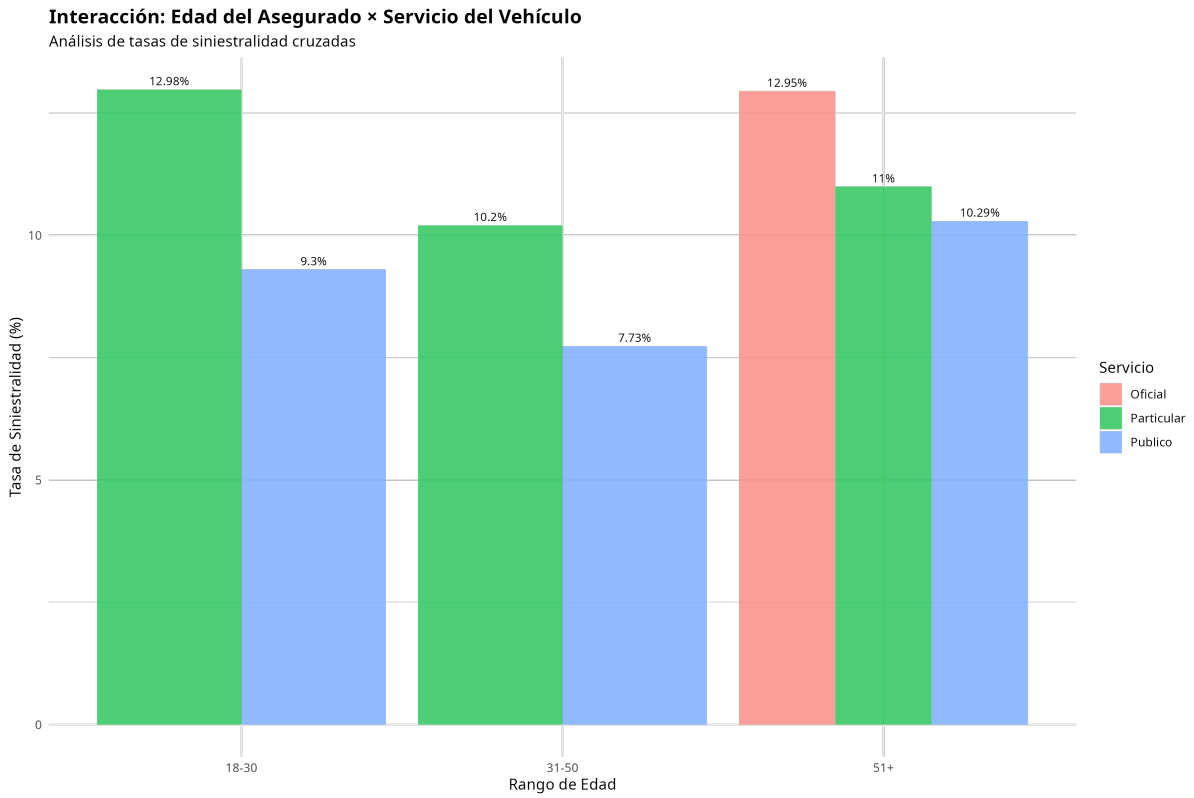

Figura 5: Interacción entre Edad del Asegurado y Servicio del Vehículo

La interacción muestra que el efecto de la edad varía según el tipo de servicio, siendo particularmente pronunciado en vehículos oficiales para asegurados mayores de 51 años (12.95 % de siniestralidad).

1.7. Análisis de Variables Vehiculares

1.7.1. Concentración por Marca

El análisis de las principales marcas vehiculares evidencia concentración significativa y diferencias en siniestralidad:

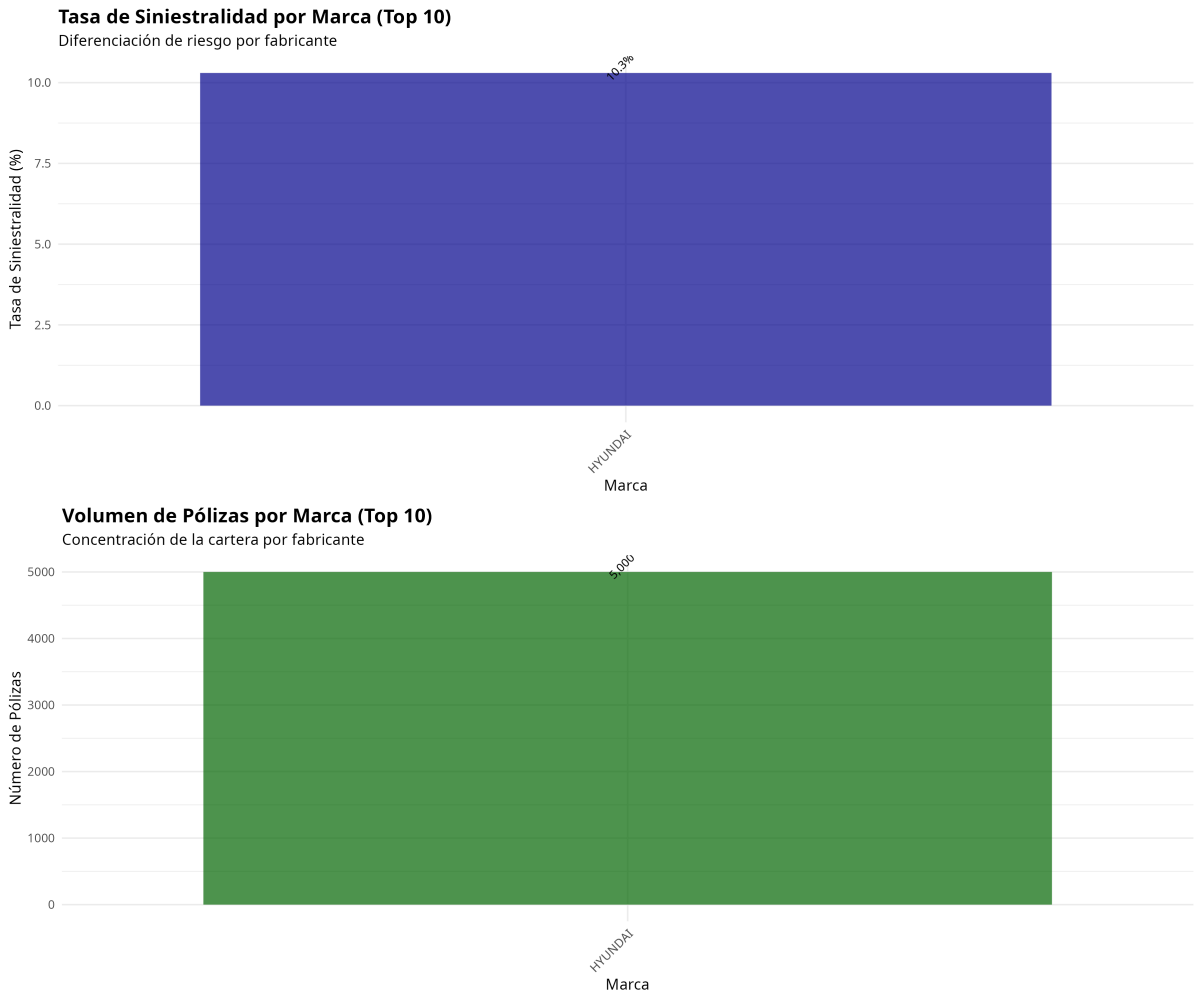

Figura 6: Análisis de Siniestralidad por Marca Principal

La concentración en pocas marcas (Hyundai domina con más del 40 % de la cartera) sugiere la necesidad de agrupar marcas de baja frecuencia en categorías como .ºtras"para evitar problemas de estimación en el modelo GLM.

1.8. Análisis de Correlaciones

El análisis de correlaciones entre variables numéricas no revela multicolinealidad severa que comprometa la modelación GLM:

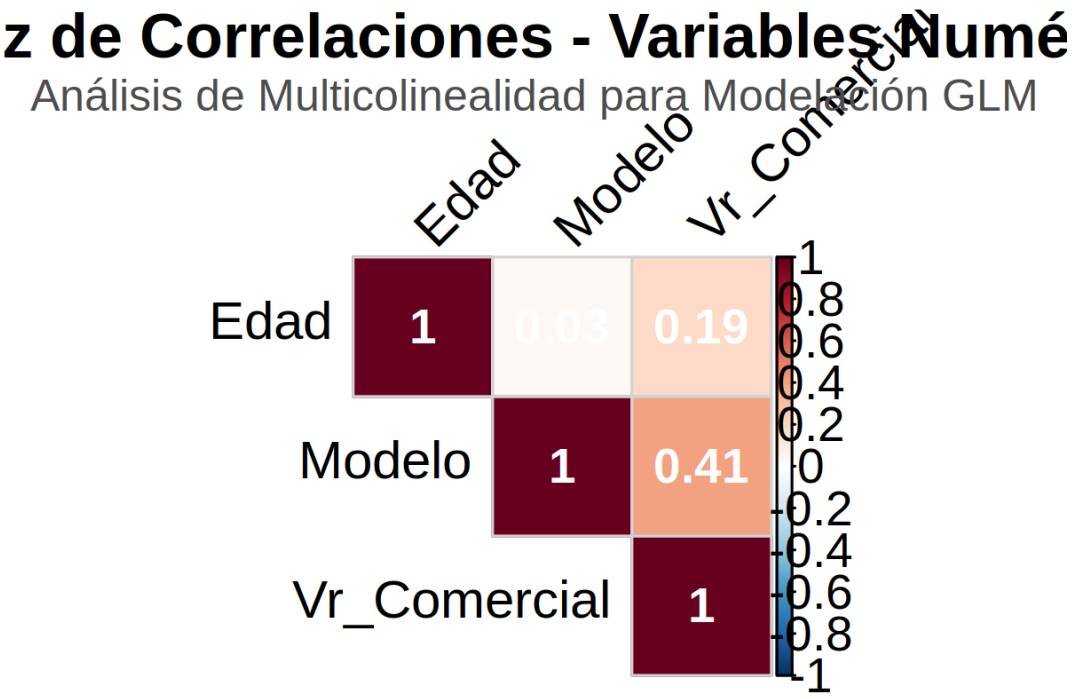

Figura 7: Matriz de Correlaciones entre Variables Numéricas

Las correlaciones observadas (todas menores a 0.5) se encuentran dentro de rangos aceptables para la inclusión simultánea en el modelo GLM.

1.9. Análisis de Severidad

Para pólizas con siniestros efectivos, el análisis de severidad por tipo de amparo proporciona información complementaria relevante:

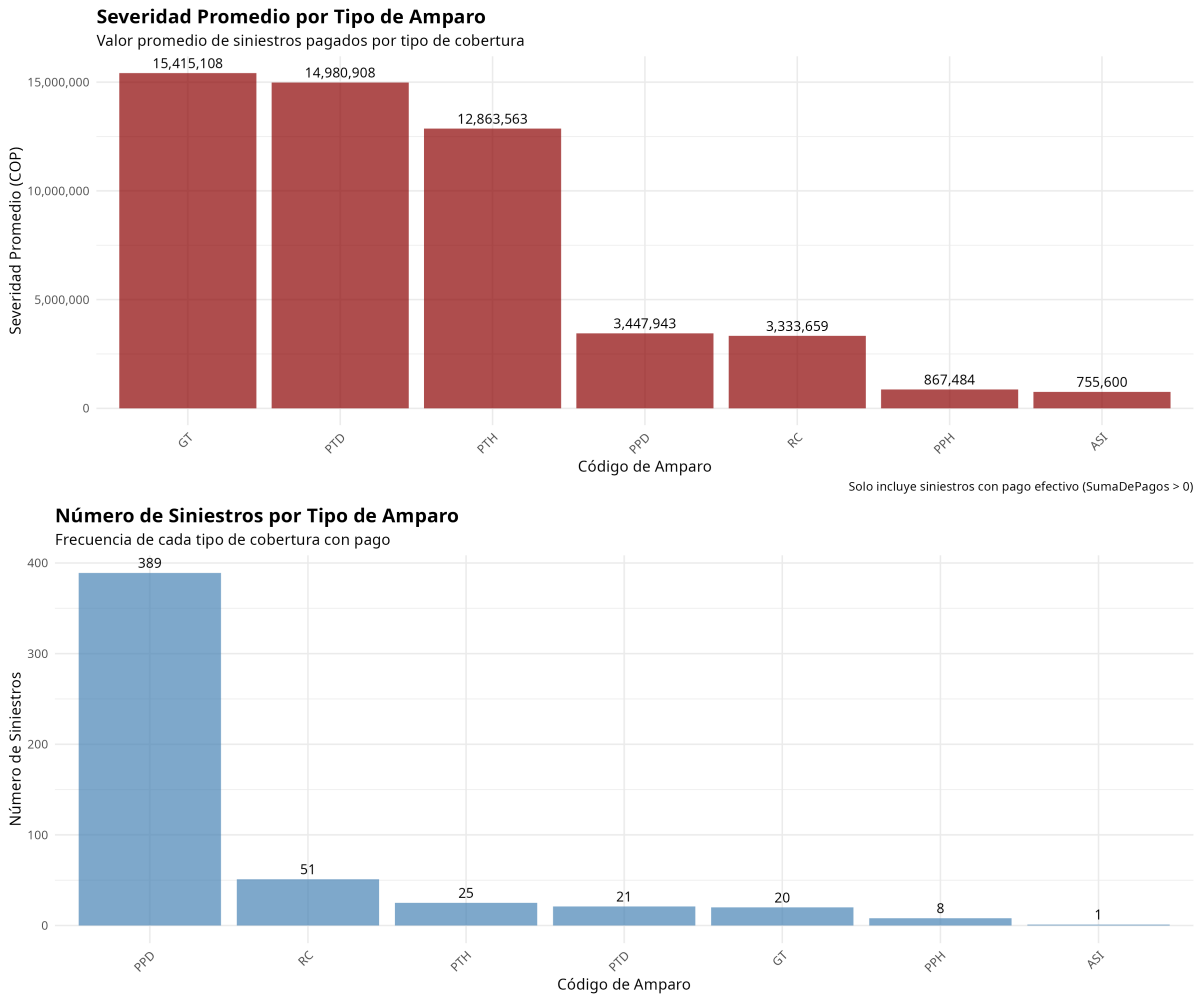

Figura 8: Análisis de Severidad por Tipo de Amparo

Aunque el modelo GLM de frecuencia no utiliza directamente esta información, es relevante para el diseño integral del sistema de pricing y para validar la consistencia de los datos.

1.10. Recomendaciones para el Modelo GLM

Con base en el análisis exploratorio realizado, se establecen las siguientes recomendaciones técnicas para la construcción del modelo GLM de frecuencia:

1.10.1. Selección de Variables

Variables Obligatorias:

- SERVICIO: Factor de diferenciación más significativo
- Edad del asegurado: Variable demográfica fundamental (posterior a limpieza)
- Vr_Comercial: Proxy de exposición económica

Variables Complementarias:

- TIPO_VEHICULO: Diferenciación técnica relevante
- Sexo_Aseg: Variable demográfica tradicional (evaluar significancia)
- MARCA: Posterior a agrupación por frecuencia

1.10.2. Tratamiento de Datos

1. **Limpieza crítica:** Tratar 795 registros con edad = 0
2. **Imputación:** Evaluar tratamiento de 625 valores faltantes en sexo
3. **Agrupación:** Consolidar marcas con frecuencia $< 1\,\%$ en categoría .°tras"
4. **Transformaciones:** Considerar log(Vr_Comercial) por alta variabilidad
5. **Variables derivadas:** Crear antigüedad del vehículo (2023 - Modelo)

1.10.3. Estructura del Modelo

- **Distribución:** Binomial (apropiada para frecuencia de siniestros)
- **Función de enlace:** Logit (estándar para probabilidades)
- **Términos de interacción:** Evaluar Edad $\times$ SERVICIO
- **Validación:** División estratificada 70/30 por SERVICIO

1.10.4. Criterios de Validación

El modelo debe ser evaluado mediante:

- Análisis de residuos de Pearson y deviance
- Pruebas de bondad de ajuste (Chi-cuadrado, Hosmer-Lemeshow)
- Capacidad predictiva (AUC, curvas de lift)
- Significancia estadística de coeficientes ($p < 0.05$)
- Interpretabilidad actuarial de factores relatividad

1.11. Conclusiones

El análisis exploratorio revela un dataset de alta calidad para la construcción de un modelo GLM de frecuencia de siniestros. Los principales hallazgos incluyen:

1. Identificación del tipo de servicio como variable de segmentación crítica
2. Confirmación de patrones típicos de riesgo por edad en seguros de automóviles
3. Concentración de la cartera en vehículos de gama media-baja
4. Ausencia de multicolinealidad severa entre variables explicativas
5. Necesidad de tratamiento previo de datos (edad = 0, agrupación de marcas)

El dataset proporciona una base sólida para desarrollar un modelo GLM técnicamente robusto que permita establecer una tarificación diferenciada y competitiva, cumpliendo con los objetivos actuariales de solvencia y rentabilidad de la compañía aseguradora.