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	**18**	**Registros: 5,000**

Dado que no tenemos un columna que indique el tipo de poliza que se tiene para cada registro se asumirá que todas las polizas en la base de datos son contra todo riezgo. Por otro lado, para facilidad y viabilidad del modelo las varaibles númericas como Modelo y Edad se van a agrupar. Las varaibles como Modelo y Edad serán agrupadas, la única variable que será siguiendo continua es la Variable de Vr_Comercial. Las variable de tipo fecha serán usadas para calcular una variable que calcule la exposición en días.

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

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. Al hacer la agrupación de esta categoría estos datos serán tratado como una categoría aparte.

Similarmente, se identificaron 308 registros (6.16 % del dataset) con valor comercial igual a 0. Estos registros comprometen el cálculo de primas basadas en valor del vehículo

EN cuanto a las variable categóricas, hay en la mayoria de ellas un exeso de grupos. Por consiguiente, el tratatamiento de los datos que resultaron relevantes como el Color será reagrupar los datos según su frecuencia; se establecío el número máximo de grupos como 4, esto con el objetico de faciliar el calculo de las tarifas.

1.3. Variables Objetivo: Frecuencia de Siniestros y Suma de Pagos

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

Cuadro 3: 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:

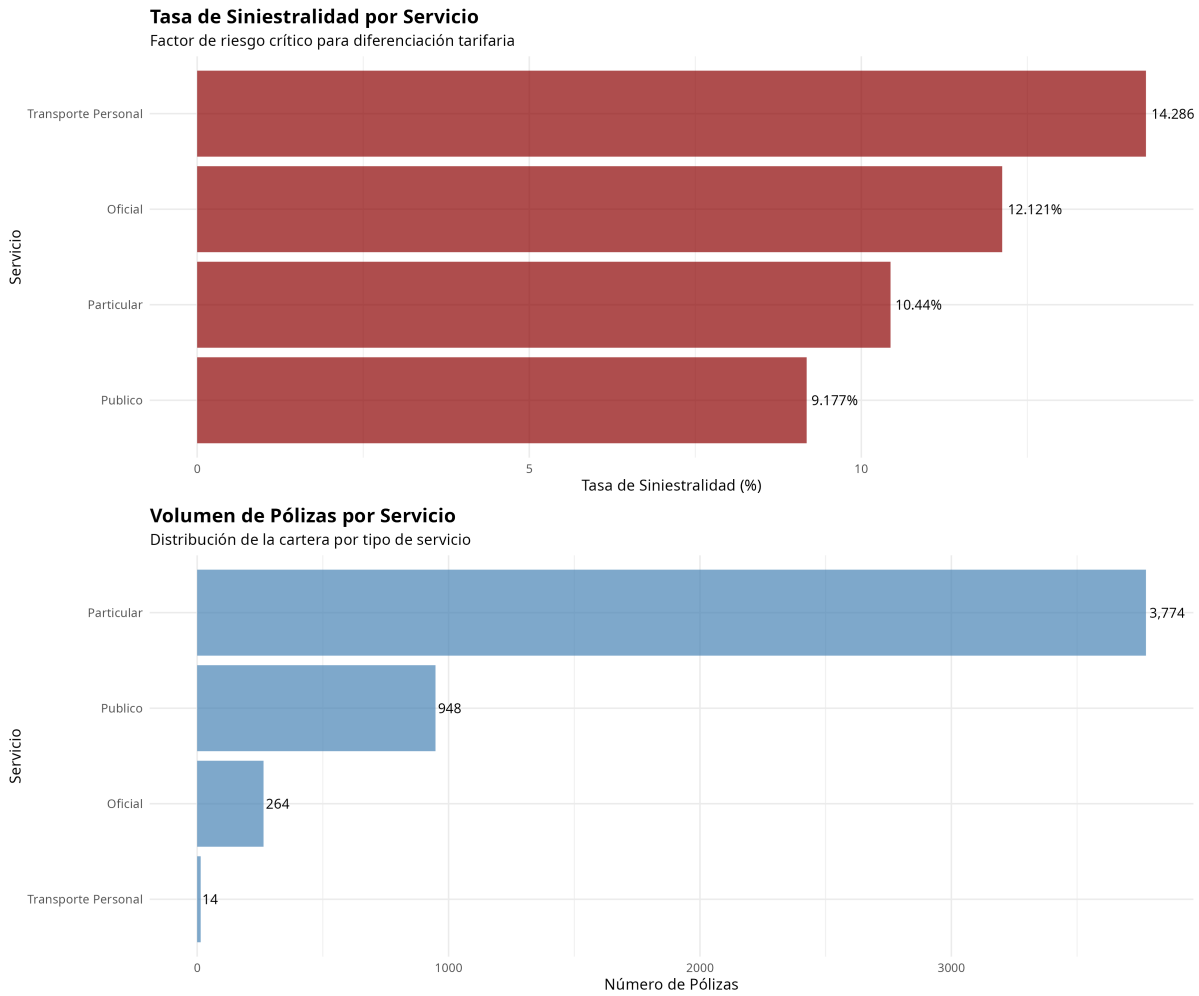

Figura 1: Análisis de Siniestralidad por Servicio del Vehículo

Cuadro 4: 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:

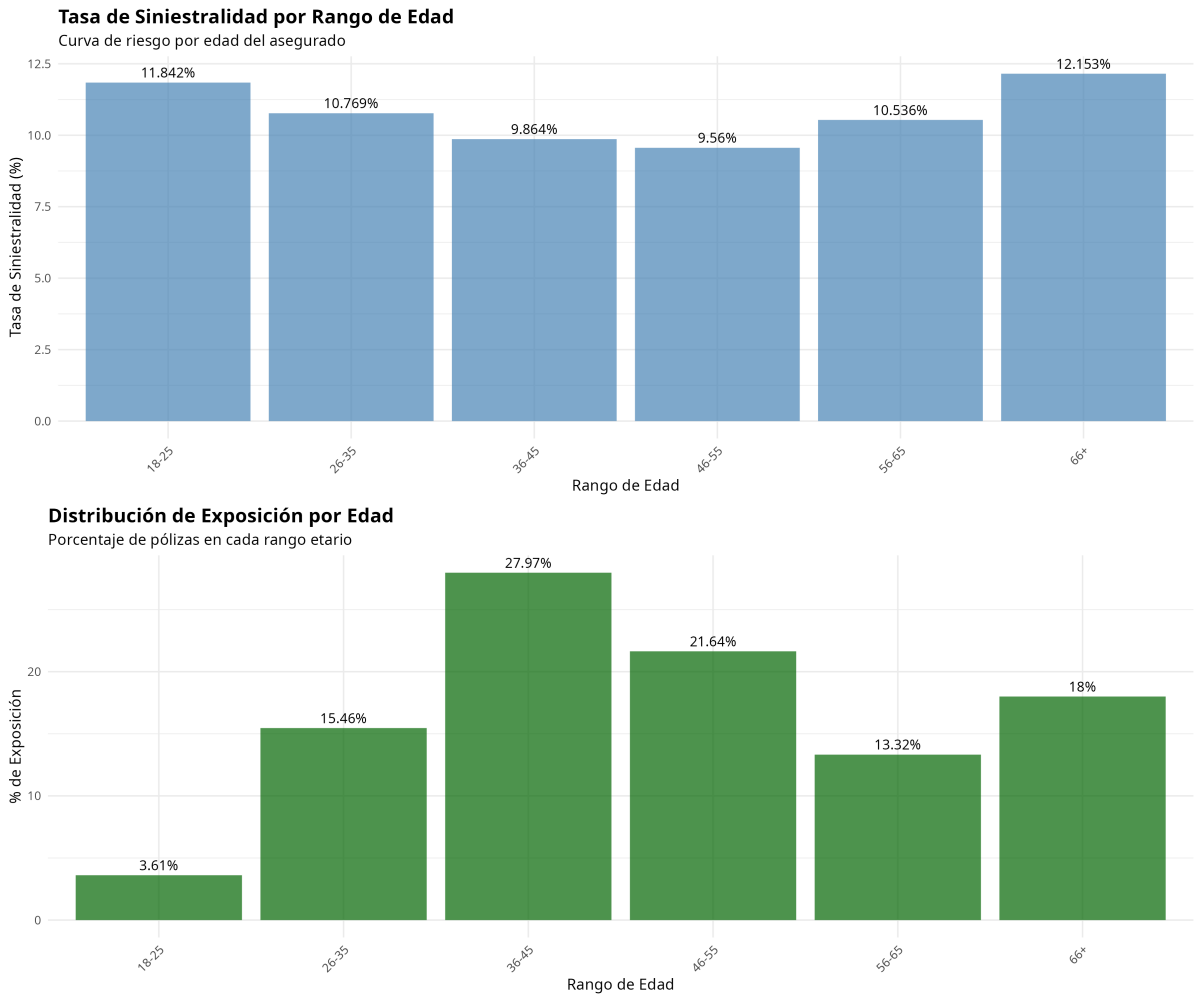

Figura 2: Análisis de Siniestralidad por Edad del Asegurado

Cuadro 5: 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:

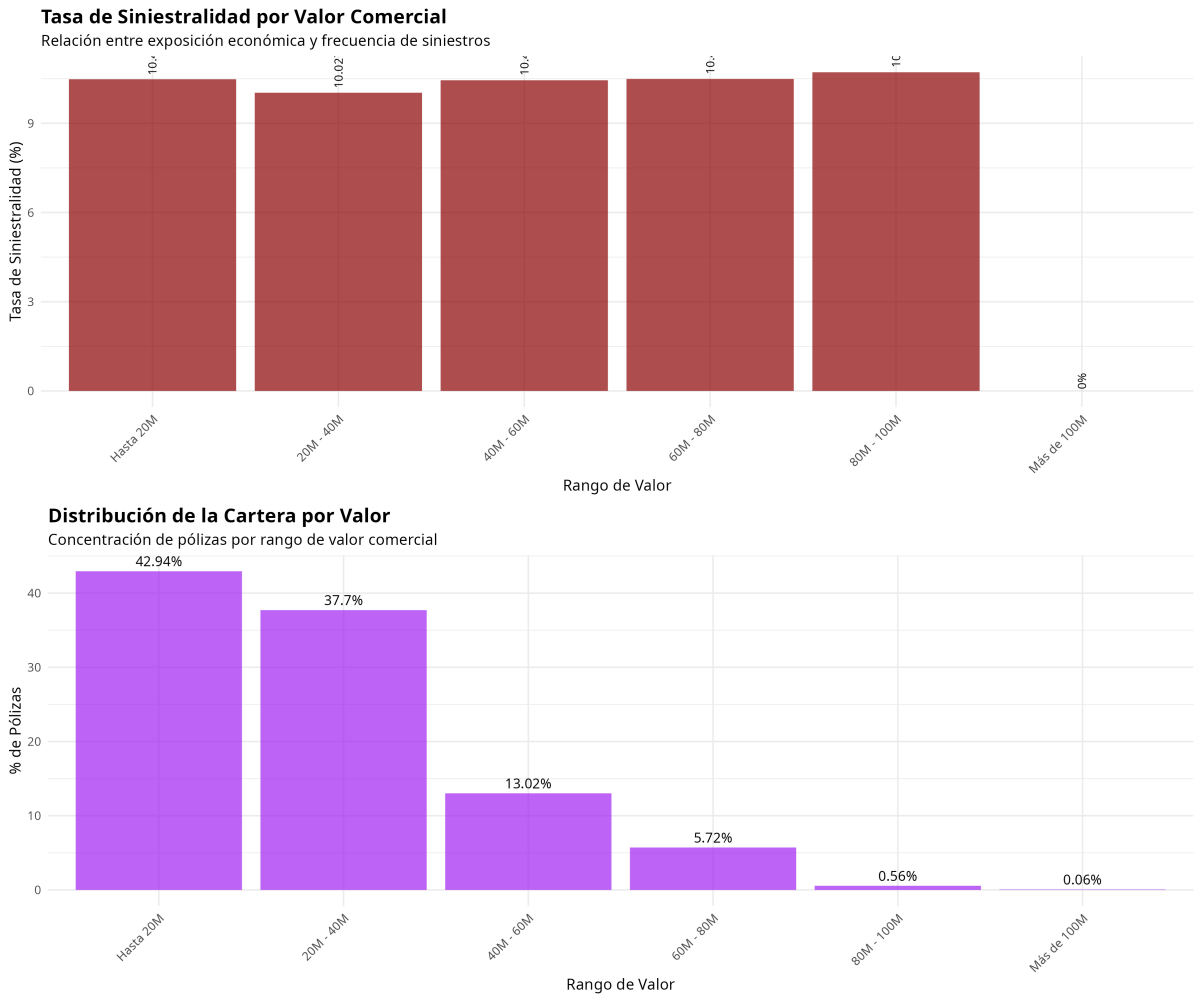

Figura 3: Distribución y Siniestralidad por Valor Comercial

Cuadro 6: 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:

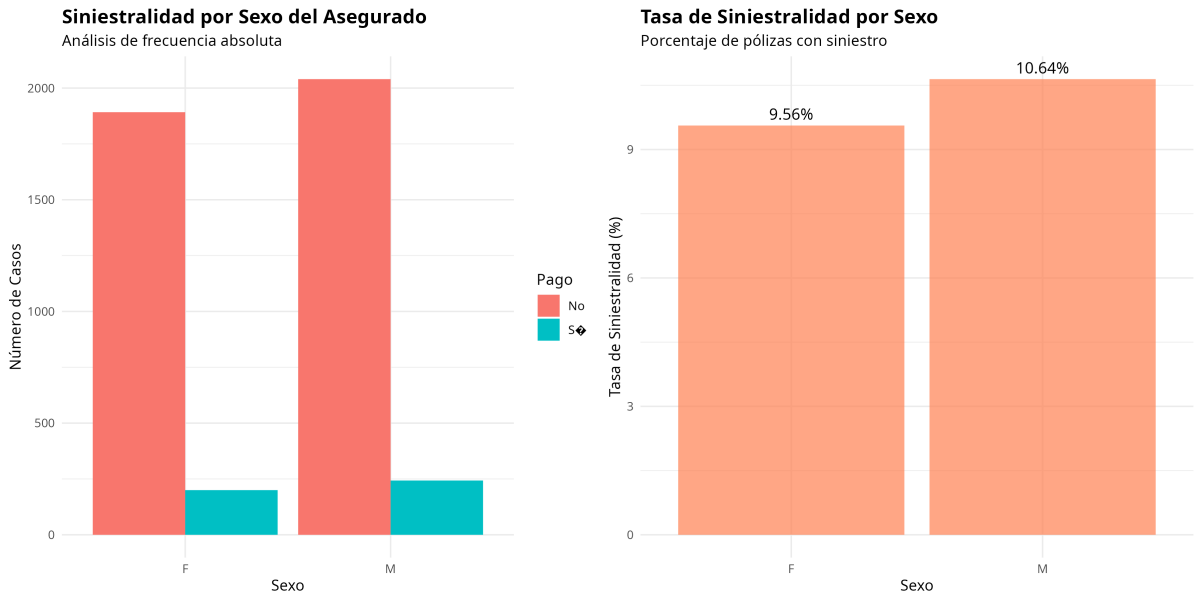

Figura 4: Siniestralidad por Sexo del Asegurado

Cuadro 7: 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 variables las categorías deben de verse por separado. A su vez, el 17 % de los datos no tienen sexo, por consiguiente serán tratados como otra categoría.

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:

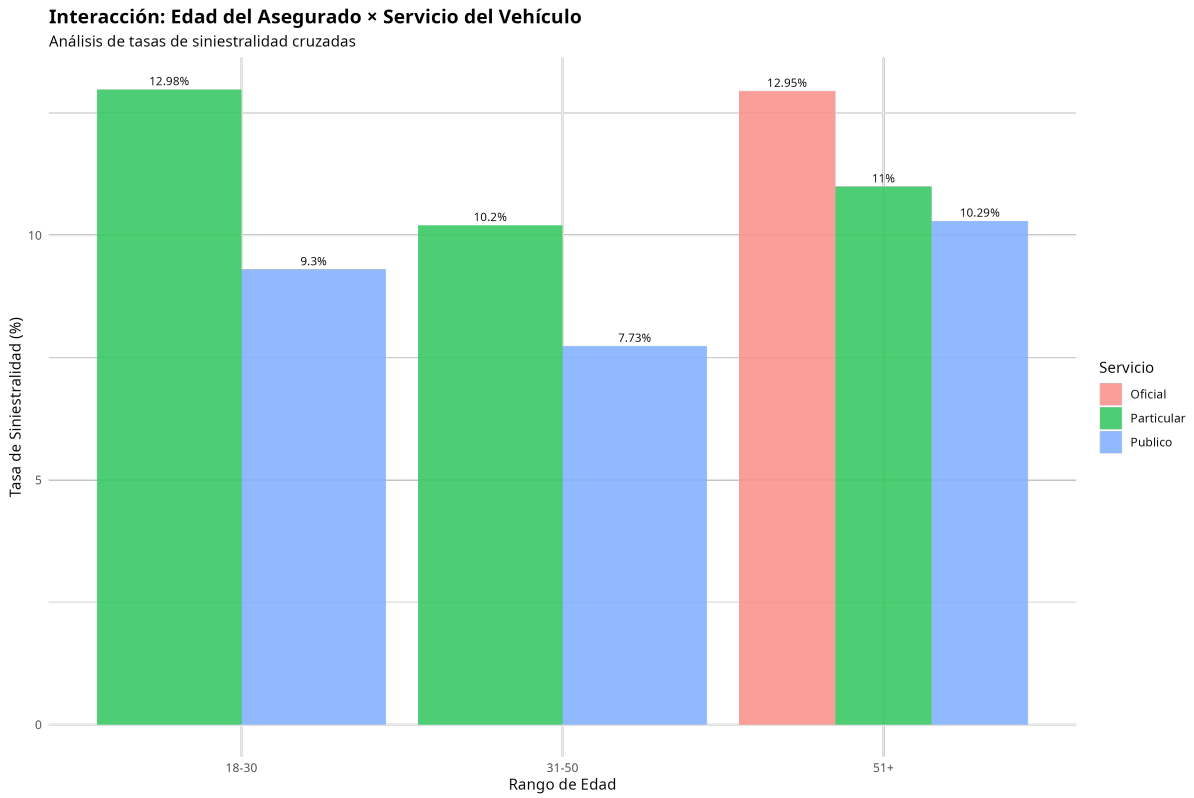

Figura 5: Interacción entre Edad del Asegurado y Servicio del Vehículo

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 Correlaciones

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

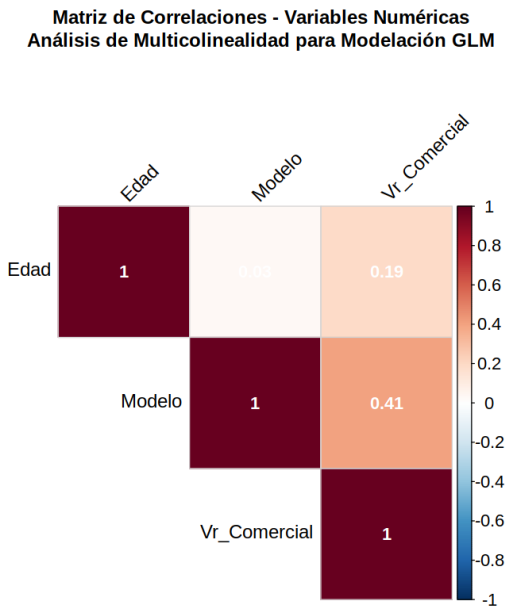

Figura 6: Matriz de Correlaciones entre Variables Numéricas

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

1.8. Análisis de Severidad

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

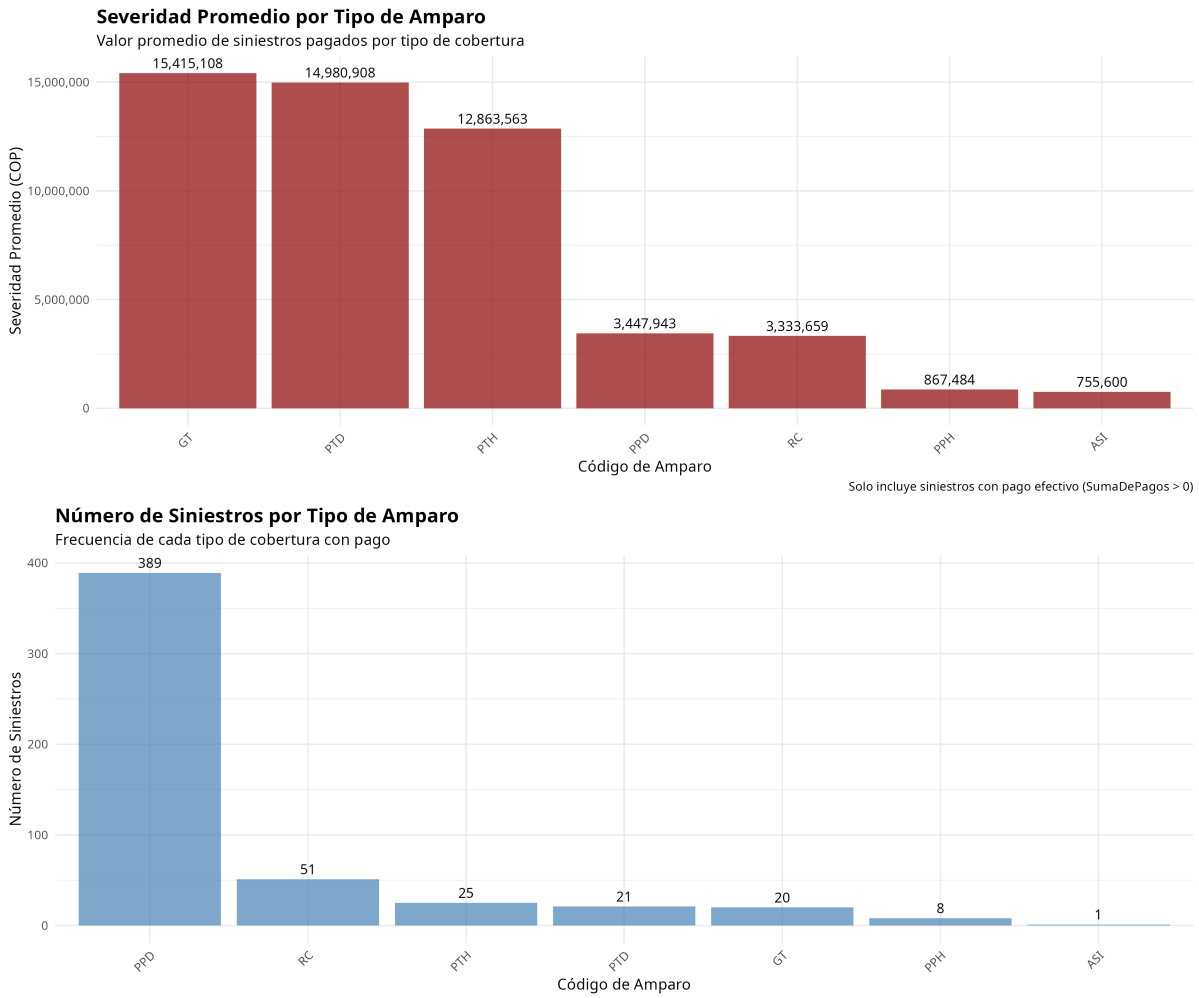

Figura 7: Análisis de Severidad por Tipo de Amparo

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.