

**UNIVERSIDAD NACIONAL DE INGENIERÍA
FACULTAD DE INGENIERÍA ECONOMICA, ESTADISTICA
Y CIENCIAS SOCIALES**



TESIS

**"PRIMA PURA SUJETO A RIESGOS DEPENDIENTES una
propuesta metodológica con múltiples coberturas"**

**PARA OBTENER EL GRADO ACADÉMICO DE MAESTRO EN
CIENCIAS EN CIENCIAS ACTUARIALES**

ELABORADO POR:

**JOSEPH GARCIA SANTIAGO
ORCID: 0009-0001-0829-7991**

ASESOR:

Ph. D. JOSÉ JAVIER CERDA HERNÁNDEZ

ORCID: 0000-0002-9297-5694

LIMA – PERÚ

2023

© 2023, Universidad Nacional de Ingeniería. Todos los derechos reservados

“El autor autoriza a la UNI a reproducir la tesis en su totalidad o en parte, con fines estrictamente académicos.”

Garcia Santiago, Joseph

jgarcias@uni.edu.pe

924932992

*Dedicado a,
mi madre Elena Santiago Peña
por su constante amor y cariño,
siempre testigo de mis logros y
que a pesar de haber estado
lejos desde mi vida universitaria
siempre me apoyó incondicionalmente,*

*mi padre Flaviano García Méndez
por creer siempre en mi capacidad,
por fortalecerme cuando creía
de no poder hacerlo, y sobretodo
por brindarme su cariño y apoyo,*

*A mi esposa Linda Quiñonez Huamanlazo
quién estuvo conmigo en los
momentos más difíciles que nos tocó
afrontar en estos últimos años.*

*mi hija Valentina
quién es mi motivo de superación
profesional.*

AGRADECIMIENTOS

Agradezco a todos mis familiares, amigos y colegas del trabajo que de alguna manera me han brindado su apoyo y han sido testigos de este logro profesional. En particular:

Agradezco a mi asesor de tesis, Phd. José Cerda , por sus enseñanzas, consejos y colaboración en lograr la presente tesis.

A mis revisores de tesis, Mg. Marco Ávila y Mg. Rafael Caparó, por su tiempo y disposición en revisar la tesis, además, de brindarme sugerencias de la misma.

Agradezco a Elizabeth Cáceres, Directora de Advisory de KPMG en Perú, por brindarme horas para desarrollar la presente tesis dentro de la jornada laboral, cuya finalidad ha sido y es, la de mantener un equipo de servicios profesionales altamente capacitado.

Agradezco a mis profesores de la Maestría en Ciencias Actuariales y el equipo de Posgrado de la Facultad de Ingeniería Económica, Estadística y Ciencias Sociales por haber llevado en adelante el desarrollo de esta maestría pionera en la formación profesionales Actuarios peruanos, además, de permitirme ser parte de la primera promoción de Actuarios que iniciamos esta maestría.

Índice general

RESUMEN	VI
ABSTRACT	VII
LISTA DE TABLAS	VIII
LISTA DE FIGURAS	IX
LISTA DE SÍMBOLOS Y SIGLAS	XI
CAPÍTULO I: Planteamiento del problema	1
1.1 Descripción de la situación problemática	1
1.2 Formulación del problema	2
1.2.1 Problema general	2
1.2.2 Problemas específicos	2
1.3 Objetivos de la investigación	2
1.3.1 Objetivo General	2
1.3.2 Objetivos Específicos	2
1.4 Justificación, alcances y limitaciones de la investigación	3
1.4.1 Justificación	3
1.4.2 Alcances	3
1.4.3 Limitaciones	3
CAPÍTULO II: Marco Teórico Conceptual	5
2.1 Antecedentes de la investigación	5
2.2 Bases teóricas de la investigación	7
2.2.1 Modelos Lineales Generalizados	7
2.2.2 Cúpulas	19
2.2.3 Familias de Cúpulas	30
2.2.4 Modelos GLM basados en Cúpulas	35
2.2.5 Construcción de modelos multivariados altamente dependientes - 2 Alternativas	42
2.3 Enfoque teórico asumido por el investigador	49
2.3.1 Prima pura sujeto a efectos de dependencia	49
2.4 Hipótesis	53
2.4.1 Hipótesis general	53

2.4.2	Hipótesis específicas	53
2.5	Variables	53
2.5.1	Robustez	53
2.5.2	Asociación	53
2.5.3	Desempeño	54
CAPÍTULO III:METODOLOGÍA		55
3.1	Tipo, nivel y diseño de la investigación	55
3.2	Población, muestra y tamaño de muestra	55
3.2.1	Población	55
3.2.2	Muestra	55
3.2.3	Tamaño de muestra	55
3.3	Técnicas de análisis e instrumentos	55
3.3.1	Técnicas de Análisis	55
3.3.2	Instrumentos	57
3.4	Operacionalización y Matriz de consistencia	63
3.4.1	Cuadro de operacionalización de variables	63
3.4.2	Matriz de consistencias	64
CAPÍTULO IV:ANALISIS Y RESULTADOS		65
4.1	Análisis descriptivo de los datos	65
4.2	Análisis de los datos	75
4.2.1	Modelos Marginales de la Frecuencia y de la Severidad	75
4.2.2	Pruebas de indicios de Dependencia	76
4.2.3	Modelos GLM Conjuntos de la Frecuencia y de la Severidad	80
4.2.4	Agregación de la Prima Pura de Riesgo por Cobertura	81
4.3	Interpretación y discusión de los resultados	87
CONCLUSIONES		90
RECOMENDACIONES		91
REFERENCIAS BIBLIOGRÁFICAS		92
ANEXOS		95
ANEXO A: ANÁLISIS UNIVARIADO		96
ANEXO B: MODELOS GLM		112
ANEXO C: INDICIO DE DEPENDENCIA Y MODELOS GLM CONJUNTOS .		137

RESUMEN

En los últimos años, la industria de seguros en el Perú ha experimentado un notable crecimiento, catalizando el desarrollo de la profesión actuarial. En el ámbito del *pricing*, es común utilizar modelos predictivos. Aunque su implementación no está completamente extendida en las compañías de seguros, es necesario contar con modelos más eficientes que nos permitan estimar primas de manera adecuada. Esto significa cumplir con nuestras obligaciones frente a los asegurados por siniestros ocurridos, cubrir los gastos asociados y, finalmente, generar un margen de utilidad para la aseguradora.

Para la estimación de la prima de riesgo es usual analizar la distribución de la frecuencia y de la severidad sobre alguna línea de negocio. La modelación para una cartera de seguros normalmente incluyen covariables asociadas al asegurado, al bien asegurado, zonas geográficas e historial de siniestralidad, es decir, la modelación de la frecuencia y de la severidad es multivariada, con base a ello los modelos lineales generalizados son una herramienta de modelación por excelencia de la frecuencia y de la severidad en función de covariables. Sin embargo, en la práctica estos riesgos se suelen asumir independientes entre sí, lo que nos conlleva a una sobre-estimación de la prima pura de riesgo.

En objetivo principal del presente trabajo, es incluir el efecto de la dependencia entre la frecuencia y la severidad en el cálculo de la prima pura. Para lograr esto utilizamos diversas familias de cópulas, cuyas propiedades teórica han demostrado un buen performance y una eficiencia en el modelamiento conjunto de riesgos con cierto grado de dependencia. Utilizamos cuatro familias de cópulas para construir un modelo conjunto de la frecuencia y de la severidad en presencia de covariables, y lo aplicamos a una línea de negocio específica de alta frecuencia, como lo son los seguros vehiculares. Además, los datos utilizados en el presente trabajo provienen de un repositorio estadístico con datos reales de otro país, pero será suficiente para entender y modelar la dependencia empírica observada entre la frecuencia y la severidad.

Finalmente, el principal aporte de este trabajo es mostrar una propuesta metodológica para el cálculo de la prima de riesgo, considerando múltiples coberturas en un entorno donde la frecuencia y la severidad son dependientes. Al final veremos que existe cierta sobre estimación de la prima de riesgo por limitarse al supuesto de independencia.

ABSTRACT

The insurance industry has boomed in recent years in Peru, and along with it the development of the actuarial profession. In the field of Pricing it is usual to use predictive models, although their application is not 100% implemented in insurance companies, it is necessary to have more efficient models that allow us to estimate sufficient premiums, that is to say, that allow us to meet our obligations in the face of claims, cover the expenses associated with the sale and finally generate a profit margin for the insurer. For the estimation of the risk premium it is usual to analyze the distribution of frequency and severity over some line of business. The modeling for an insurance portfolio normally includes covariates associated with the insured, the insured property, geographical areas and claims history, i.e., the modeling of frequency and severity is multivariate, based on which generalized linear models are a modeling tool par excellence for frequency and severity as a function of covariates. However, in practice these risks are usually assumed to be independent of each other, which leads to an overestimation of the pure risk premium.

The present work aims to include the effect of the dependence between frequency and severity, to achieve this we will use copulas functions, whose properties have demonstrated their efficiency in the joint modeling of risks with a certain degree of dependence. We will use four families of copulas to build a joint model of frequency and severity in the presence of covariates and applied on a high frequency line of business such as vehicle insurance, in addition, the information used comes from a statistical repository with real data from another country, but it will be enough to understand the relationship between frequency and severity.

Finally, the contribution of this work is to show a methodological proposal for the calculation of the risk premium, considering multiple coverages in an environment where frequency and severity are dependent. At the end we will see that there is some overestimation of the risk premium by limiting it to the assumption of independence.

LISTA DE TABLAS

Tabla N° 2.1	Distribuciones de la familia exponencial y sus parámetros	9
Tabla N° 2.2	Funciones enlaces comunes	11
Tabla N° 2.3	Devianza para distribuciones de la familia exponencial	17
Tabla N° 2.4	Datos del Cuarteto de Anscombe	25
Tabla N° 2.5	Relación entre el parámetro θ y τ de Kendall	36
Tabla N° 2.6	Primera derivada parcial	37
Tabla N° 2.7	Parámetros del modelo de distribución conjunta	38
Tabla N° 2.8	Trasformación no restricta de θ	40
Tabla N° 2.9	Expresiones de v por cada familia de cópulas	52
Tabla N° 3.1	Cuadro de operacionalización de variables	63
Tabla N° 3.2	Matriz de consistencia	64
Tabla N° 4.1	Variable: Tipo de póliza	66
Tabla N° 4.2	Variable: Antiguedad del vehículo	67
Tabla N° 4.3	Variable: Canal de venta	68
Tabla N° 4.4	Variable: Uso del vehículo	69
Tabla N° 4.5	Variable: Clase del vehículo	70
Tabla N° 4.6	Variable: Tipo de vehículo	70
Tabla N° 4.7	Variable: Marca del vehículo y por clase	71
Tabla N° 4.8	Variable: Zona demográfica	72
Tabla N° 4.9	Variable: Género del asegurado	73
Tabla N° 4.10	Variable: Tipo de Persona	73
Tabla N° 4.11	Variable: Edad del asegurado	74
Tabla N° 4.12	Modelo Marginal de Frecuencia y Severidad - Pérdidas Parciales	76
Tabla N° 4.13	Modelo Marginal de Frecuencia y Severidad - Pérdidas Totales .	77
Tabla N° 4.14	Modelo Marginal de Frecuencia y Severidad - Responsabilidad Civil .	77
Tabla N° 4.15	Modelo Marginal de Frecuencia y Severidad - Asistencias	77
Tabla N° 4.16	Pérdida Parcial Log-likelihood de los modelos de regresión . . .	80
Tabla N° 4.17	Pérdida Total Log-likelihood de los modelos de regresión	80
Tabla N° 4.18	Responsabilidad Civil Log-likelihood de los modelos de regresión	80
Tabla N° 4.19	Asistencias Log-likelihood de los modelos de regresión	80
Tabla N° 4.20	Modelo de Regresión GLM Conjunto - Pérdidas Parciales	81
Tabla N° 4.21	Modelo de Regresión GLM Conjunto - Pérdidas Totales	82
Tabla N° 4.22	Modelo de Regresión GLM Conjunto - Responsabilidad Civil . .	82
Tabla N° 4.23	Modelo de Regresión GLM Conjunto - Asistencias	82
Tabla N° 4.24	Grados de dependencia bivariado - Tau de Kendall	85

LISTA DE FIGURAS

Figura N° 2.1	Superficie máxima Log-Verosimilitud	14
Figura N° 2.2	Distribución Chi-Cuadrado	18
Figura N° 2.3	Descripción de como se utiliza el Teorema de Sklar para calcular la distribución conjunta de un vector aleatorio.	21
Figura N° 2.4	Función de densidad conjunta	22
Figura N° 2.5	Gráficos del Cuarteto de Anscombe	26
Figura N° 2.6	Chi-plot para variables independientes con $p = 0.95$	29
Figura N° 2.7	Chi-plot para variables dependientes con $p = 0.95$	29
Figura N° 2.8	Gráficos K-PLOT	30
Figura N° 2.9	Cópula Gaussiana ($\rho = 0.8$)	31
Figura N° 2.10	Cópula Student($\tau = 0.5, \nu = 4$)	32
Figura N° 2.11	Cópula Clayton($\theta = 0.5$)	33
Figura N° 2.12	Cópula de Frank($\theta = 0.5$)	34
Figura N° 2.13	Cópula de Gumbel($\theta = 0.5$)	35
Figura N° 2.14	D-vine cópula: 5 variables, 4 árboles y 10 pair-cópulas	45
Figura N° 2.15	C-vine cópula: 5 variables, 4 árboles y 10 pair-cópulas	45
Figura N° 2.16	Ilustración del árbol $\tau = \{\emptyset, (1), (1, 1), (1, 2), (2)\}$	46
Figura N° 2.17	Una ilustración de un modelo de agregación de 4 dimensiones .	48
Figura N° 2.18	Densidad conjunta de la Frecuencia y Severidad	51
Figura N° 2.19	Distribución de probabilidad de riesgos agregados	52
Figura N° 3.1	Medida de exposición de una póliza en un periodo de tiempo .	56
Figura N° 3.2	Algoritmo para la construcción del Simple Quantile Plot . .	58
Figura N° 3.3	Ejemplo del <i>loss ratio chart</i>	59
Figura N° 3.4	Índice de Gini para un Modelo de Prima Pura	60
Figura N° 4.1	Frecuencia y Severidad por cobertura y periodo de análisis .	65
Figura N° 4.2	Matriz de correlación - Pérdida Parcial	75
Figura N° 4.3	Robustez de los modelos de GLM Marginales - Índice de GINI .	78
Figura N° 4.4	Indicador de dependencia entre la Frecuencia y la Severidad por Cobertura	78
Figura N° 4.5	Chi-Plots de la Frecuencia y Severidad por cobertura	79
Figura N° 4.6	K-Plots de la Frecuencia y Severidad por cobertura	79
Figura N° 4.7	Distribución de la Prima Pura de Riesgo	83
Figura N° 4.8	Ajuste a una distribución Gamma - Estimación de parámetros .	84
Figura N° 4.9	Formas de agregación de las primas por cobertura	84
Figura N° 4.10	Esquema de agregación de riesgos por cobertura	85
Figura N° 4.11	Distribución de la Prima de Riesgo Bivariada	86

Figura N° 4.12	Distribución de la Prima de Riesgo Multivariada Pérdida Parcial + Pérdida Total + Responsabilidad Civil + Asistencias	87
Figura N° 4.13	Prima Pura de Riesgo: Independencia vs. Dependencia	89
Figura N° .1	Agrupación de la variable Antiguedad del vehículo	96
Figura N° .2	Agrupación de la Marca Vehículo - Pérdida Parcial - Livianos . .	97
Figura N° .3	Agrupación de la Marca Vehículo - Pérdida Total - Livianos . .	98
Figura N° .4	Agrupación de la Marca Vehículo - Responsabilidad Civil - Li- vianos	99
Figura N° .5	Agrupación de la Marca Vehículo - Asistencias - Livianos . . .	100
Figura N° .6	Agrupación de la Marca Vehículo - Pérdida Parcial - Pesados .	101
Figura N° .7	Agrupación de la Marca Vehículo - Pérdida Total - Pesados . .	102
Figura N° .8	Agrupación de la Marca Vehículo - Responsabilidad Civil - Pe- sados	103
Figura N° .9	Agrupación de la Marca Vehículo - Asistencias - Pesados . . .	104
Figura N° .10	Agrupación de la Marca Vehículo - Pérdida Parcial - Vehículos Menores	105
Figura N° .11	Agrupación de la Marca Vehículo - Pérdida Total - Vehículos Menores	106
Figura N° .12	Agrupación de la Marca Vehículo - Responsabilidad Civil - Vehícu- los Menores	107
Figura N° .13	Agrupación de la Marca Vehículo - Asistencias - Vehículos Me- nores	108
Figura N° .14	Edad del asegurado - Sin Agrupar	109
Figura N° .15	Edad del asegurado - Agrupado en rangos	110
Figura N° .16	Rangos de Suma Asegurada del Vehículo	111
Figura N° .17	Matriz de correlación - Pérdida Total	112
Figura N° .18	Matriz de correlación - Responsabilidad Civil	113
Figura N° .19	Matriz de correlación - Asistencias	114

LISTA DE SÍMBOLOS Y SIGLAS

SÍMBOLOS

- α : Vector de coeficientes del modelo de Severidad
 β : Vector de coeficientes del modelo de Frecuencia
 θ : Parámetro de la función Cúpula
 λ : Vector esperado de la Frecuencia
 μ : Vector esperado de la Severidad
 ϕ : Parámetro de dispersión del modelo de Severidad
 τ : Tau de Kendall - Parámetro de dependencia

SIGLAS

- PPR : Prima Pura de Riesgo
PP : Cobertura Pérdida Parcial
PT : Cobertura Pérdida Total
RC : Cobertura Responsabilidad Civil
AS : Cobertura Asistencias
GLM : Generalized Linear Model

CAPÍTULO I: Planteamiento del problema

1.1 Descripción de la situación problemática

Los cambios normativos relacionados al sector asegurador han ido exigiendo a las compañías de seguros una medición más rigurosa y detallada de los riesgos a los que están expuestos para reducir su probabilidad de insolvencia. Dentro de estos requerimientos cuantitativos es fundamental obtener modelos que incluyan de una mejor forma los riesgos a los que se está expuesto, para obtener una mejor estimación de la reserva técnica, que es una provisión que debe tener la compañía de seguros para afrontar los pagos de los siniestros. En tal sentido, lo anterior se traduce en un cálculo de primas de riesgo más exactas utilizando modelos estadísticos y actuariales más sofisticados y reales.

Sin embargo, para el cálculo de la prima pura dentro de la teoría clásica de riesgo es común asumir por simplicidad que las distribuciones de la frecuencia y de la severidad son independientes. Como consecuencia, la prima pura de riesgo es el producto de la frecuencia promedio y la severidad promedio. Este supuesto utilizado por muchos practitioners en la industria de seguros, a menudo injustificado y rebatido por evidencia empírica ([Czado, Kastenmeier, Brechmann, y Min, 2012](#); [Lee, Park, y Ahn, 2018](#); [Ohlsson y Johansson, 2010](#); [P. Song, Li, y Yuan, 2009](#)), limita el cálculo de la prima de riesgo generando una sobre estimación. Es por ello que encontrar modelos donde la dependencia entre la frecuencia y la severidad sean considerados pueden reducir aún más la varianza del estimador de la prima de riesgo, generando un estimador más robusto.

Otro problema abordado en la presente investigación es que cuando las primas de riesgo son calculadas por coberturas, es usual en la práctica agregarlas usando una suma simple de riesgos para obtener la prima de riesgo total. Esto genera nuevamente una sobre estimación de la cuantificación del riesgo, pudiendo generar una prima menos competitiva y atractiva para los clientes. Sin embargo, la evidencia empírica muestra que en el momento que ocurre un siniestro suelen activarse más de una cobertura ([Anderson et al., 2007](#)). Esto muestra que existe un nivel de dependencia que debe considerarse cuando las primas por cobertura son agregadas. En el presente trabajo también abordamos este problema que existe en la industria de seguros, incorporando cópulas para modelar la dependencia y estimar mejor el riesgo agregado.

Ante lo expuesto, se propone construir modelos más eficientes para la estimación de la prima pura de riesgo, considerando los efectos de la dependencia que existe entre riesgos que se quieren coberturar. Para la modelación conjunta de la frecuencia y severidad por cobertura proponemos utilizar modelos lineales generalizados basados en cópulas mixtas, y para la agregación de las primas de riesgo por cobertura construimos una cópula multivariada que logre agregarlos a un solo valor, es decir, llegar a la prima pura de riesgo total.

1.2 Formulación del problema

1.2.1 Problema general

- ¿Es posible construir una metodología de cálculo de la prima pura de riesgo sin suponer la independencia entre la frecuencia y severidad? ¿Es posible agregar diferentes riesgos mediante cópulas por cobertura de tal forma que nos permita obtener estimaciones confiables y más justas de la prima de riesgo para los asegurados frente al supuesto usual de independencia?

1.2.2 Problemas específicos

- ¿Es posible determinar un modelo lineal generalizado conjunto de la frecuencia y de la severidad para estimar la prima pura de riesgo de una determinada cobertura?
- ¿Cómo podemos medir el grado de dependencia entre la frecuencia y la severidad, e incorporarla en el modelo?
- ¿Es posible determinar una propuesta metodológica para el cálculo de la prima total mediante agregación de los riesgos asociados de cada cobertura mediante funciones cópulas?
- ¿Cómo podemos medir el grado de dependencia entre las primas por cobertura para el cálculo de la prima total?

1.3 Objetivos de la investigación

1.3.1 Objetivo General

- Proponer una propuesta metodológica del cálculo de la prima pura de riesgo sujeto a dependencia entre los riesgos de frecuencia y severidad, y agregados mediante cópulas por cobertura que nos permita estimaciones confiables y de acuerdo al riesgo específico de cada asegurado frente supuesto usual de independencia.

1.3.2 Objetivos Específicos

- Determinar un modelo de regresión generalizado conjunto de la frecuencia y de la severidad para estimar la prima pura de riesgo de una determinada cobertura.
- Medir el grado de dependencia entre la frecuencia y la severidad.
- Determinar una propuesta metodológica de cálculo de la prima total mediante agregación de los riesgos asociados de cada cobertura mediante funciones cópulas.
- Medir el grado de dependencia entre las primas por cobertura para el cálculo de la prima total o prima agregada.

1.4 Justificación, alcances y limitaciones de la investigación

1.4.1 Justificación

La importancia de la presente investigación radica en que actualmente en la industria de seguros la mayoría de las empresas aseguradoras, y sobretodo en mercados emergentes como el nuestro, donde la profesión de actuaria es cada vez más demandada, las técnicas suelen limitarse al uso de conceptos teóricos básicos y de uso muy práctico que no necesariamente son correctas o que asumen demasiados supuestos que no son realistas, generando un sesgo en las estimaciones de la prima de riesgo. El presente trabajo pretende reducir esta brecha al mostrar que se pueden aplicar conceptos teóricos más sofisticados de matemática actuarial, estadística y algoritmos de optimización para la construcción de modelos más eficientes para el cálculo de la prima de riesgo, desde el punto de vista estadístico. Además, la metodología descrita en el presente trabajo permite obtener una prima de acuerdo al riesgo específico de cada asegurado, y a la vez atractiva para los asegurados que adquieren un producto de seguro, puesto que al considerar la dependencia entre la severidad, la frecuencia y los riesgos cubiertos el método para el cálculo de prima propuesto no usa la misma información que puede estar contenida en diferentes tipos de riesgos, generando una menor prima respecto al supuesto de independencia, como lo demuestra diversos trabajos en la literatura ([Anderson et al., 2007; Czado et al., 2012; Garrido, Genest, y Schulz, 2016; Krämer, Brechmann, Silvestrini, y Czado, 2013; Lee, Park, y Ahn, 2019; P. Song et al., 2009; P. X.-K. Song, 2000](#)).

1.4.2 Alcances

- Para el modelamiento de la dependencia utilizaremos solo las cópulas de Gauss, Clayton, Gumbel y Frank; siendo estas familias de cópulas las de mayor uso.
- Si bien pueden existir diversas coberturas en producto de seguros, el presente trabajo solo abarcará las coberturas de Pérdida Parcial, Pérdida Total, Responsabilidad Civil y Asistencias en seguros vehiculares.
- Se mostrará el efecto de la inclusión del efecto de la dependencia en la distribución de la prima pura de riesgo y de la pérdida total observada.

1.4.3 Limitaciones

- Nuestro análisis no se incluye algún esquema de recargos por gastos y márgenes de utilidad sobre la prima pura de riesgo total, que es objeto del presente trabajo. Los recargos obedecen a estrategias comerciales de acuerdo a sus canales de venta, y no son de naturaleza estocástica.
- La propuesta del presente trabajo de investigación es sólo aplicable a productos de seguros diferentes a vida.

- Los modelos predictivos del presente trabajo no se implementarán en sistemas informáticos empresariales de tal manera que nos permita observar en tiempo real su desempeño. Sin embargo, realizamos simulaciones mediante plantillas de hojas de cálculo para ver su aplicabilidad y desempeño.
- EL costo computacional de la implementación del algoritmo desarrollado en este trabajo es alto, y requiere de cierta expertise para su implementación.

CAPÍTULO II: Marco Teórico Conceptual

2.1 Antecedentes de la investigación

La Teoría de Riesgo Colectivo, introducida en (Lundberg, 1903), proporciona una marco científicamente aceptado por la academia y la industria para el planteamiento y resolución de los principales problemas y desafíos que presentan las actividades de la industria aseguradora, donde la incertidumbre es una de las principales características que la definen (Kaas, Goovaerts, Dhaene, y Denuit, 2008). La siniestralidad es la variable aleatoria característica del negocio asegurador, cuyo análisis parte en entender sus componentes básicos: El número de siniestros y la cuantía de un siniestro. Bajo el supuesto de independencia de estas componentes o variables aleatorias, el valor esperado de la siniestralidad o más conocida como la prima pura de riesgo se obtiene como el producto de ambos (McNeil, Frey, y Embrechts, 2005; Ohlsson y Johansson, 2010; Sikov y Cerdá-Hernández, 2021).

Sin embargo, dentro de las características de los riesgos asegurables por las compañías de seguros, es que estos deben ser homogéneos, con la finalidad de obtener una prima suficiente y competitiva que garantice la solvencia de la empresa. Es en ese sentido que los modelos lineales generalizados son herramientas muy utilizadas cuando se requiere incluir covariables o factores de riesgo que nos permitan segmentar la cartera total de una empresa de seguros en subsegmentos homogéneos, donde la cuantía de la prima pura de riesgo será proporcional a su grado de siniestralidad. El primer ejemplo aplicado a la industria de seguros donde se usan los GLM, específicamente a seguros de motores, fue presentado en (McCullagh y Nelder, 1989), y fue popularizado en la industria de seguros después de los pioneros trabajos de Renshaw (Renshaw, 1994) y Brockman y Wright (Brockman y Wright, 1992), donde se reconoce el potencial de los GLM como una herramienta de modelado integral para la frecuencia y la severidad en presencia de covariables. Además, estos modelos prestan especial atención a una diversa variedad de distribuciones disponibles, y a las técnicas de estimación de parámetros mediante conceptos de *cuasi-likelihood*.

Desde la introducción de los GLM para el análisis de carteras de seguros hecha por Renshaw, Brockman y Wright, los modelos lineales generalizados son y siguen siendo una de las técnicas estadísticas más utilizadas para la construcción de modelos multivariados para la frecuencia y para la severidad (Anderson et al., 2007; Czado et al., 2012; Lee et al., 2019). En la práctica, la evidencia empírica muestra que la frecuencia y la severidad son a menudo dependientes, y es por ello que existe la necesidad de incluir en los modelos la asociación potencial o dependencia existente (Czado et al., 2012; Lee et al., 2018; Ohlsson y Johansson, 2010; P. Song et al., 2009). Una forma de estudiar y capturar la dependencia entre variables aleatorias fue introducida en los trabajos pioneros de Sklar (Sklar, 1959), y específicamente aplicado a la industria de seguros en (Garrido et al., 2016), donde los autores consideran el número de siniestros como una covariable adicional en el modelamiento del costo medio, es

decir, lo que los autores proponen en dicho trabajo es construir una distribución condicional de la severidad respecto de la frecuencia, y que en términos prácticos, la prima pura de riesgo resulta como el producto de tres elementos: la media de la distribución marginal de la frecuencia, la media de la distribución modificada de la severidad y un factor de corrección por dependencia. En la misma dirección, (Frees, Lee, y Yang, 2016) construyen primero la distribución conjunta de la severidad y la frecuencia usando la distribución de la frecuencia y la distribución condicional de la severidad al centrarse en el uso de una cópula para modelar la dependencia entre estos.

Desde que se comprobó que el coeficiente de correlación lineal de Pearson no es un instrumento adecuado para medir relaciones de dependencia entre riesgos, el uso de las cópulas juega un papel importante para modelar dependencia, puesto que superan muchas de las limitaciones del coeficiente de correlación de Pearson. Las cópulas fueron presentadas y caracterizadas para variables continuas por Sklar (Sklar, 1959), quién resolvió algunos de los problemas formulados por M. Fréchet sobre la relación entre una función de distribución de probabilidad multidimensional y sus marginales de menor dimensión. Actualmente, la teoría de cópulas se ha convertido en una poderosa herramienta de modelado multivariado en muchos campos donde la dependencia entre varias variables aleatorias marginales, continuas o discretas, es de gran interés, y para los cuales la suposición de normalidad multivariada puede ser cuestionable, supuesto utilizado para usar el coeficiente de correlación de Pearson. Es por ello que existen investigaciones donde modelan la frecuencia y la severidad de forma conjunta mediante modelos lineales generalizados basados en cópula mixta aplicados a datos de Salud (de Leon y Wu, 2011) y a datos de automóviles (Czado et al., 2012; Frees et al., 2016). Otros trabajos interesantes, basados en las investigaciones anteriores, pero ampliando su aplicación sobre un mayor número de familias de cópulas se pueden encontrar en (Czado et al., 2012) para estimar la pérdida total, y en (Kholifah, Lestari, y Devila, 2019) para estimar la prima pura de riesgo. Ambos trabajos proponen algoritmos de optimización en la estimación de los parámetros y procedimientos adecuados para la selección óptima de la cópula.

Finalmente, si bien en este trabajo nosotros utilizamos diferentes modelos ya desarrollados en la literatura de cópulas para realizar el modelamiento conjunto de la frecuencia y de la severidad para el cálculo de la prima pura de riesgo; en la presente tesis también proponemos una metodología para resolver un problema actual que se tiene en la industria de seguros en el Perú: calcular la prima pura de riesgo total de un seguro que ofrece varias coberturas simultáneamente. Este problema se observa empíricamente al activarse varias coberturas de forma simultánea ante un mismo siniestro. Sabemos que las distribuciones de la frecuencia y de la severidad se relacionan de manera distinta por cada cobertura, obteniendo una prima pura de riesgo por cada una de ellas, pero el problema es que actualmente la industria está tarifando un seguro con varias coberturas haciendo una suma simple de las primas obtenidas, sin incluir el grado de dependencia que pueda existir entre ellas. En este trabajo propone-

mos una metodología para tarifar este tipo de seguro incluyendo la dependencia entre ellas y construyendo una función conjunta multivariada (Aas, Czado, Frigessi, y Bakken, 2009) que logré agregar dichas primas incluyendo la dependencia entre ellas. De esta manera obtendremos primas precisas que garanticen la suficiencia de la compañía, con primas competitivas para el mercado de seguros.

2.2 Bases teóricas de la investigación

2.2.1 Modelos Lineales Generalizados

Como la teoría de Modelos Lineales Generalizados es una teoría bien estudiada actualmente en estadística, para el desarrollo de la base metodológica de este trabajo seguiremos De Jong, Heller, et al. (2008), Rao, Shalabh, Toutenburg, y Heumann (2010), Ohlsson y Johansson (2010).

Estos modelos son utilizados para calcular y cuantificar la relación entre la variable de respuesta y las variables explicativas. Estos modelos difieren de un modelo de regresión clásicos en dos importantes aspectos:

1. La distribución de la variable de respuesta es elegida de una familia de distribuciones, denominada *familia exponencial*. Por lo tanto, la distribución de la variable de respuesta no necesita ser Normal o cercana a ello, o puede ser explícitamente no Normal.
2. Una transformación de la media de la variable respuesta es linealmente relacionada a las variables explicativas.

La consecuencia de permitir que la variable respuesta sea parte de la familia exponencial es que la respuesta puede ser, y suele ser, heterocedástica. Así, la varianza variará con la media que, a su vez, puede variar con las variables explicativas. Esto contrasta con el supuesto homocedástico de los modelos de regresión con residuo normal. Los modelos lineales generalizados son ampliamente utilizados en la industria de seguros, puesto que el supuesto de normalidad no aplica a ese tipo de datos. Por ejemplo, la distribución de las pérdidas generadas por los siniestros no sigue una distribución Normal (para más detalles ver Rao et al. (2010), Ohlsson y Johansson (2010)).

2.2.1.1 Modelos lineales generalizados

Los Modelos lineales generalizados (GLM por sus siglas en inglés) son una generalización de los clásicos modelos lineales de análisis de regresión y análisis de varianza. Los parámetros de estos modelos son estimados de acuerdo al principio de mínimos cuadrados y son optimizados de acuerdo a la teoría de dispersión mínima (Rao et al., 2010).

Según McCullagh y Nelder (1989), un modelo GLM consiste de tres componentes:

1. La *componente aleatoria*, que especifica la distribución de probabilidad de la variable de interés,
2. El *componente sistemático*, que especifica una función lineal de las variables explicativas,
3. La *función de enlace*, que describe una relación funcional entre el componente sistemático y la esperanza de la componente aleatoria.

Las tres componentes pueden ser especificadas de forma compacta de la siguiente manera:

La componente aleatoria consiste de una variable respuesta y , donde la distribución pertenece a la familia exponencial, es decir, la función de densidad de probabilidad de cada observación de la variable y tiene la forma:

$$f(y_i) = c(y_i, \phi) \exp\left[\frac{y_i \theta - a(\theta)}{\phi}\right], \quad g(\mu) = x' \beta. \quad (2.1)$$

donde θ y ϕ son los parámetros. El parámetro θ es llamado parámetro canónico y ϕ es el parámetro de dispersión. Cualquier función de probabilidad que pueden ser escrito como en (2.1) se dicen que pertenecen a la familia exponencial. En términos de $a(\theta)$ podemos calcular la media y la varianza de y :

$$E(y) = \dot{a}(\theta), \quad \text{Var}(y) = \phi \ddot{a}(\theta),$$

donde $\dot{a}(\theta)$ y $\ddot{a}(\theta)$ son la primera y segunda derivada de $a(\theta)$ respecto de θ , respectivamente. En la Tabla 2.1 se muestra las diferentes formas de θ y $a(\theta)$ para algunas distribuciones que pertenecen a las distribuciones de la familia exponencial.

Aquí no mostramos la forma de la expresión $c(y, \phi)$ dado que no suele ser de interés en la mayoría de situaciones y aplicaciones. En el caso del parámetro de dispersión de (2.1) suele expresarse como ϕ/w , donde w es un ponderador. Esta forma es apropiada cuando tenemos datos agrupados. La segunda ecuación de (2.1) nos dice que una transformación de la media de la variable y , $g(\mu)$, es igual a una combinación lineal de las variables explicativas contenidas en el vector x , donde g es la función de enlace. Algunas observaciones sobre el modelo GLM son:

- La elección de la función $a(\theta)$ determina la distribución de la variable de respuesta.
- La elección de enlace $g(\mu)$ determina cómo la media está relacionada con las variables explicativas x . En el modelo de regresión lineal general, la relación entre la media de y y las variables explicativas es $\mu = x' \beta$. En cambio, en los modelos GLM, esto se generaliza como $g(\mu) = x' \beta$, donde g es una función monótona y diferenciable (tal como la función logaritmo o la función raíz cuadrada).

TABLA N° 2.1: Distribuciones de la familia exponencial y sus parámetros

Función de Distribución	θ	$a(\theta)$	ϕ	$E(y)$	$V(\mu) = \frac{Var(y)}{\phi}$
Binomial(n, π)	$\ln \frac{\pi}{1-\pi}$	$n \ln(1+e^\theta)$	1	$n\pi$	$n\pi(1-\pi)$
Poisson(μ)	$\ln \mu$	e^θ	1	μ	μ
Normal(μ, σ^2)	μ	$\frac{1}{2}\theta^2$	σ^2	μ	1
Gamma(μ, ν)	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{\nu}$	μ	μ^2
Gamma Inversa(μ, σ^2)	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2	μ	μ^3
Binomial Negativa(μ, κ)	$\ln \frac{\kappa\mu}{1+\kappa\mu}$	$-\frac{1}{\kappa} \ln(1-\kappa e^\theta)$	1	μ	$\mu(1+\kappa\mu)$

- De la ecuación (2.1) observamos que dado x , el parámetro μ es determinado a través de la relación $g(\mu)$. De igual forma dado μ , el parámetro θ es determinado a través de $a(\theta) = \mu$. Finalmente dado θ, y es determinada como un extracto de la densidad exponencial especificada en $a(\theta)$.
- Las observaciones del vector y se asumen independientes.

2.2.1.2 Pasos en el modelado lineal generalizado

Dada una variable de respuesta y , la construcción de un GLM consta de los siguientes pasos:

1. Elija una distribución de respuesta $f(y)$ y, por lo tanto, elija $a(\theta)$ en (2.1). La distribución de la variable de respuesta se adapta a la situación dada.
2. Seleccione una función enlace $g(\mu)$. Esta elección a veces se simplifica eligiendo el llamado enlace “canónico” correspondiente a cada distribución de la variable de respuesta.
3. Elija las variables explicativas x en términos de las cuales se va a modelar $g(\mu)$. Se aplican consideraciones similares a las del modelo de regresión lineal ordinaria.
4. Recoja las observaciones y_1, \dots, y_n sobre la respuesta y y los valores correspondientes x_1, \dots, x_n sobre las variables explicativas x . Se supone que las observaciones sucesivas son independientes, es decir, la muestra se considerará como una muestra aleatoria simple de la población.

5. Ajuste el modelo estimando los parámetros contenidos en β y, si se desconoce, también estimadmos el parámetro ϕ . Para estimar los parámetros usamos el método de máxima verosimilitud.
6. Dada la estimación de β , genere predicciones (o valores ajustados) de y para diferentes configuraciones de x , y examine qué tan bien se ajusta el modelo examinando la desviación de los valores ajustados de los valores reales, así como otros diagnósticos del modelo. También se usará el valor estimado de β para ver si las variables explicativas dadas son o no importantes para determinar μ .

Los pasos adicionales en comparación con el modelo de regresión lineal ordinario son elegir la función $a(\theta)$ (lo que implica la distribución de la variable de respuesta) y la función enlace $g(\mu)$. La elección $a(\theta)$ está guiada por la naturaleza de la variable de respuesta. La elección de la función enlace viene sugerida por la forma funcional de la relación entre la respuesta y las variables explicativas.

En ejemplo reales, los pasos anteriores rara vez proceden en dicha secuencia, especialmente con datos de seguros. Los datos a menudo se recopilan antes de la especificación de un modelo. Por ejemplo, en la industria de los seguros, todas las pólizas de un tipo particular son colectadas y analizadas y modeladas posteriormente. Es probable que la exploración inicial de los datos sugiera diferentes modelos y diferentes distribuciones para la variable respuesta. Los ajustes iniciales a menudo van seguidos de refinamientos adicionales de ajustes, en los que algunas de las variables explicativas pueden descartarse o transformarse. Los datos pueden ser maquillados juiciosamente por la exclusión de varios casos, o la incorporación de diferentes efectos.

2.2.1.3 Función de enlace y enlace canónico

Las funciones de enlace $g(\mu)$ más utilizadas en las aplicaciones a datos reales son mostradas en la Tabla 2.2. Con la excepción de la función de enlace para el modelo logit, el resto de funciones de enlace son de la forma $g(\mu) = \mu^p$, siendo el caso logarítmico el límite de $(y^p - 1)/p$ cuando $p \rightarrow 0$.

Si $g(\mu) = \theta$, entonces g se denomina el enlace canónico correspondiente a $a(\theta)$. En este caso tenemos que θ es una combinación lineal de las variables explicativas, es decir, $\theta = x' \beta$. Elegir el enlace canónico correspondiente a una distribución de respuesta f simplifica computacionalmente el proceso de estimación de los parámetros, aunque en la actualidad con la capacidad computacional moderna que se tiene esto ya no es una consideración primordial que se toma en cuenta al momento de usar el modelo GLM. En la Tabla 2.2 se muestran las distribuciones de respuesta para las cuales los enlaces comúnmente utilizados son canónicos. Las constantes en θ generalmente se omiten del enlace canónico.

TABLA N° 2.2: Funciones enlaces comunes

Función enlace	$g(\mu)$	Distribución
Identidad	μ	Normal
Logaritmo	$\ln\mu$	Poisson
Potencia	μ^p	Gamma ($p = -1$)
Raíz cuadrada	$\sqrt{\mu}$	
Logit	$\ln \frac{\mu}{1-\mu}$	Binomial

2.2.1.4 Variable *Offset*

Cuando la variable respuesta es una variable de conteo, como el número de siniestros en una cartera de seguros, o el número de muertes en un grupo de riesgo, requieren de un factor de corrección para el número de expuesto al riesgo n . Si μ es la media de la variable de conteo y , entonces la tasa de ocurrencia o frecuencia μ/n de nuestra variable de interés se expresa como:

$$g\left(\frac{\mu}{n}\right) = x' \beta$$

Cuando g es la función logarítmica, esto se expresa como:

$$\ln\left(\frac{\mu}{n}\right) = x' \beta \Rightarrow \ln\mu = \ln(n) + x' \beta$$

En la expresión de arriba n es llamado exposición y $\ln(n)$ es denominada la variable *offset*. El *offset* se puede interpretar como otra variable de x en la regresión, colocando un coeficiente β igual a la unidad.

Incluyendo el *offset*, y tiene un valor esperado directamente proporcional a la exposición:

$$\mu = n e^{x' \beta}$$

Los *offsets* se utilizan para corregir el tamaño del grupo o los diferentes períodos de tiempo de observación o de exposición.

2.2.1.5 Estimación por Máxima Verosimilitud

La estimación por máxima verosimilitud (MLE por sus siglas en inglés) de los parámetros β y ϕ son obtenidos de la maximización de la función *log-likelihood* que se define de la siguiente manera:

$$l(\beta, \phi) = \sum_{i=1}^n \ln f(y_i; \beta, \phi) = \sum_{i=1}^n \left[\ln c(y_i, \phi) + \frac{y_i \theta_i - a(\theta_i)}{\phi} \right] \quad (2.2)$$

el cual asume que las observaciones de la variable respuesta y_i son independientes y distribuidas con una función de densidad que pertenece a la familia exponencial.

Para la estimación del parámetro β_j mediante MLE tenemos que derivar la función $l(\beta, \phi)$ respecto de β_j :

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j}$$

donde

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - a(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi}, \quad \frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i} x_{ij}.$$

Aquí $\eta_i = x_i' \beta$ y x_{ij} es el componente i -ésimo de x_j . Establecer $\partial l / \partial \beta_j = 0$ produce las condiciones de primer orden para la maximización de la función de verosimilitud:

$$\sum_{i=1}^n \frac{\partial \theta_i}{\partial \eta_i} x_{ij} (y_i - \mu_i) = 0 \iff X' D(y - \mu) = 0 \quad (2.3)$$

Donde D es la matriz diagonal cuyos valores en la diagonal son $\partial \theta_i / \partial \eta_i$,

$$\left(\frac{\partial \theta_i}{\partial \eta_i} \right)^{-1} = \frac{\partial \eta_i}{\partial \theta_i} = \frac{\eta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \theta_i} = \dot{g}(\mu_i) \ddot{a}(\theta_i) = \dot{g}(\mu_i) V(\mu_i)$$

Por lo tanto, D es una matriz diagonal con elementos en la diagonal dados por $(\dot{g}(\mu_i) V(\mu_i))^{-1}$. Las ecuaciones en (2.3) son denominadas a menudo ecuaciones de estimación para β . Note que β está implícito en estas ecuaciones, trabajando implícitamente a través de μ y D . Definimos las matrices diagonales G y W con entradas diagonal dadas por $\dot{g}(\mu_i)$ y $[(\dot{g}(\mu_i))^2 V(\mu_i)]^{-1}$, respectivamente. Entonces $D = WG$ y la ecuación (2.3) es equivalente a:

$$X' WG(y - \mu) = 0 \quad (2.4)$$

Relación con los mínimos cuadrados ponderados: Utilizando una proximación mediante series de Taylor tenemos la siguiente aproximación

$$g(y_i) \approx g(\mu_i) + \dot{g}(\mu_i)(y_i - \mu_i) \Rightarrow g(y) \approx g(\mu) + G(y - \mu) \quad (2.5)$$

donde $g(y)$ es un vector con elementos $g(y_i)$. De forma similarmente tenemos la aproximación para $g(\mu)$. Reordenando y sustituyendo $\mu = X\beta$ obtenemos $G(y - \mu) \approx g(y) - X\beta$. Sustituyendo esta relación dentro de la ecuación (2.4) se obtiene la ecuación de estimación aproximada

$$X' W g(y) - X' W X \beta \approx 0 \rightarrow \hat{\beta} \approx (X' W X)^{-1} X' W g(y) \quad (2.6)$$

De la relación de la izquierda en (2.5), $(\dot{g}(\mu_i))^2 V(\mu_i)$ es la aproximación de primer orden

de la varianza de $g(y_i)$, además del factor ϕ . Así, la resolución de la ecuación de estimación de (2.4) corresponde, aproximadamente, a la regresión ponderada de las respuestas transformadas $g(y_i)$ con pesos proporcionales a las varianzas.

Con un enlace Identidad $g(y_i) = y_i$, la ecuación (2.5) es exacta. Entonces,

$$\hat{\beta} = (X'WX)^{-1}X'Wy$$

donde W tiene diagonal con elementos $1/V(\mu_i)$. Por lo tanto, con una función enlace Identidad y $V(\mu_i)$ independiente de μ_i , la estimación máxima verosimil para β es el estimador mínimos cuadrados ponderados (para más detalles ver [McCullagh y Nelder \(1989\)](#) y [Rao et al. \(2010\)](#)).

Método de Newton-Raphson: Obtener soluciones explícitas o cerradas de las condiciones de primer orden (2.3) para maximizar la función de verosimilitud son usualmente difíciles, excepto para casos muy particulares como la Normal con enlace Identidad. Una forma de resolver este problema es recurrir a métodos numéricos. El método más utilizado para resolver las ecuaciones que aparecen del método de MV es el método de Newton, donde se supone que la primera y segunda derivada de la función a maximizar pueden evaluarse fácilmente en cada punto. Usando estas derivadas, se construye una aproximación cuadrática a la función de verosimilitud. El maximizador resultante del método de Newton se usa luego para derivar una nueva ecuación cuadrática que, a su vez, se maximiza. A menudo se encuentra que esta secuencia de maximizadores aproximados converge rápidamente al máximo buscado ([Burden, Faires, y Burden, 2015](#)).

Para simplificar la discusión supongamos que ϕ es conocido y denotemos por $l(\beta)$ el *log-likelihood* como una función del vector parámetro desconocido β . Si β contiene un único parámetro entonces la aproximación de la serie de Taylor en cualquier punto β es dado por

$$l(\beta + \delta) \approx l(\beta) + \dot{l}(\beta)\delta + \frac{\delta^2}{2}\ddot{l}(\beta)$$

Derivando respecto de δ e igualando a cero se obtiene

$$\dot{l} + \delta\ddot{l}(\beta) = 0 \quad \Rightarrow \quad \delta = -[\ddot{l}(\beta)]^{-1}\dot{l}(\beta)$$

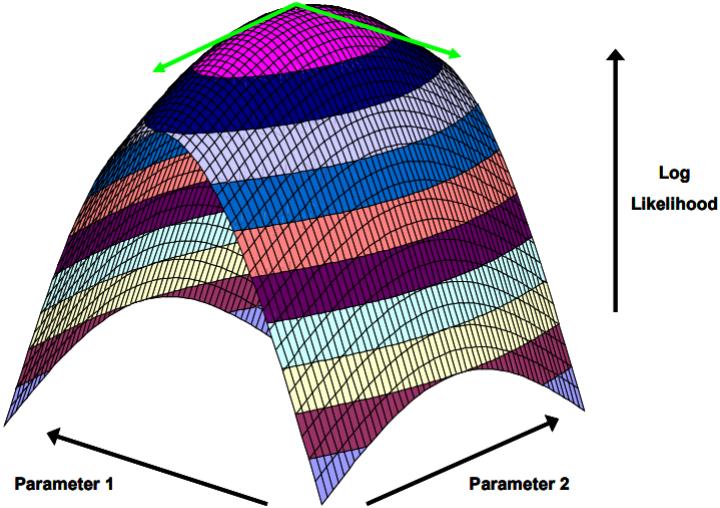
Con β dado y δ especificado, un punto más alto parece ser $\beta - \dot{l}(\beta)/\ddot{l}(\beta)$. Denotando $\beta^{(m)}$ como el valor de β en la iteración m , la ecuación de actualización es

$$\beta^{(m+1)} = \beta^{(m)} - [\ddot{l}(\beta^{(m)})]^{-1}\dot{l}(\beta^{(m)}) \tag{2.7}$$

La sucesión de esta ecuación converge a β . Además, como se mencionó anteriormente, esta sucesión converge rápidamente. El método puede ser adaptado cuando β es un vector. En este caso se realiza una aproximación cuadrática a la superficie de $l(\beta)$, y es esta superficie

cuadrática la que se maximiza

FIGURA N° 2.1: Superficie máxima Log-Verosimilitud



Fuente: Imagen tomada de [Anderson et al. \(2007\)](#)

La ecuación de actualización vista en (2.7) con la inversa interpretada como inversión de matriz, y donde $\dot{l}(\beta)$ es el vector de derivadas parciales $\partial l / \partial \beta_j$. El vector $\dot{l}(\beta)$ es llamado el *vector score*, y $\ddot{l}(\beta)$, la matriz de derivadas parciales cruzadas $\partial^2 l / (\partial \beta_j \partial \beta_k)$, la matriz *Hessiana*. La condición para un máximo es que la matriz *Hessiana* sea definida no positiva, es decir, $-(\ddot{l}(\beta))$ no sea definida negativa. El procedimiento de evaluar repetidamente el *score* y la matriz de *Hessiana* para actualizar la estimación en (2.7), se denomina iteración de Newton-Raphson.

Fisher Scoring: Fisher sugirió reemplazar $\ddot{l}(\beta)$ en (2.7) por su esperado $E(\ddot{l}(\beta))$. También demostró que dado β , $E(\ddot{l}(\beta)) = -E(\dot{l}(\beta)\dot{l}'(\beta))$. Estos resultados son ciertos para todos los *log-likelihoods*, no solo para aquellas de la familia exponencial. La matriz $-E(\ddot{l}(\beta))$ es llamada la matriz de información de Fisher. Para los modelos GLM la matriz de información es

$$E[\dot{l}(\beta)\dot{l}'(\beta)] = \phi^{-2} X' D E[(y - \mu)(y - \mu)'] DX = \phi^{-1} X' W X \quad (2.8)$$

Por lo tanto, sustituyendo $\ddot{l}(\beta)$ por su esperado en (2.7) y utilizando la ecuación (2.4) para $\dot{l}(\beta)$ obtenemos

$$\beta^{(m+1)} = \beta^{(m)} + [X' W X]^{-1} X' W G(y - \mu)$$

Esta ecuación suele reescribirse como

$$\beta^{(m+1)} = [X' W X]^{-1} X' W [X\beta^{(m)} + G(y - \mu)], \quad (2.9)$$

donde la expresión entre corchetes a la derecha se denomina "variable dependiente local". La ecuación (2.9) es similar a la regresión de mínimos cuadrados ponderados. La variable de-

pendiente local es calculado reemplazando μ por el estimado $\mu^{(m)}$ donde $g(\mu^{(m)}) = X\beta^{(m)}$. Calculando $\beta^{(m+1)}$, se requiere $V(\mu)$, la función varianza de la distribución y $\dot{g}(\mu)$, la derivada de la función enlace. El parámetro de dispersión ϕ no es requerido.

Para n grande, la inversa de la matriz de información de Fisher es aproximadamente la matriz de covarianza de $\hat{\beta}$. Además, los MLE son asintóticamente insesgados y, otra vez para muestras grandes, aproximadamente normal. Por lo tanto,

$$\hat{\beta} \approx N \left[\beta, \phi(X'WX)^{-1} \right].$$

Este resultado asintótico es la base para la prueba de significancia de Wald, que será discutido en la sección 2.2.1.8. En la práctica W es evaluado en $\hat{\beta}$.

Estimación de la dispersión: El parámetro de dispersión ϕ puede ser estimado por MLE o por el método de los momentos. MLE implica la solución iterativa de la ecuación $\partial l / \partial \phi = 0$, el cual es diferente para cada distribución de la variable de respuesta. La estimación de ϕ es generalmente considerada periférica al GLM.

2.2.1.6 Predicción e intervalos de confianza

Dada las variables explicativas x , el valor estimado de la media de y es $\hat{\mu}$, donde $g(\hat{\mu}) = x'\hat{\beta}$. Por ejemplo, con un enlace logarítmico:

$$\ln\hat{\mu} = x'\hat{\beta} \Rightarrow \hat{\mu} = e^{x'\hat{\beta}}.$$

Un intervalo de confianza alrededor del valor estimado es utilizado para indicar precisión. El cálculo del intervalo de confianza requiere de la distribución muestral de $\hat{\mu}$. La varianza del predictor lineal $x'\hat{\beta}$ es

$$Var(x'\hat{\beta}) = \phi x'(X'WX)^{-1}x$$

Por lo tanto, una aproximación del intervalo de confianza para la media, denotado por (μ_l, μ_u) , viene dado por

$$g(\mu_l) = x'\hat{\beta} - z\sqrt{\phi x'(X'WX)^{-1}x} , \quad g(\mu_u) = x'\hat{\beta} + z\sqrt{\phi x'(X'WX)^{-1}x}$$

donde z es un punto apropiado de la distribución normal estándar $N(0, 1)$. El intervalo de confianza es exacto solo en el caso de un enlace identidad y una variable de respuesta Normal. En otros casos se trata solo de una aproximación, cuya precisión mejora al aumentar el tamaño de la muestra. El parámetro de dispersión ϕ se reemplaza por una estimación, lo que genera más errores de aproximación. El estimador $\hat{\mu}$ es insesgado cuando se utiliza el enlace identidad. Para otros enlaces resulta ser sesgado. Para ilustrar el problema, usamos

el enlace logarítmico y asumimos que $\hat{\beta}$ es aproximadamente normal. Entonces,

$$E(e^{x' \hat{\beta}}) = \exp \left[x' \hat{\beta} + \frac{1}{2} \text{Var}(x' \hat{\beta}) \right] \neq e^{x' \hat{\beta}} = \mu$$

El sesgo se incrementa con la varianza $\text{Var}(x' \hat{\beta})$. Además, para muestras grandes se sabe que $\text{Var}(\hat{\beta})$ es pequeño y se espera que el sesgo sea insignificante ([McCullagh y Nelder, 1989](#); [Rao et al., 2010](#)). El intervalo (μ_l, μ_u) es un intervalo de confianza para la media de y dado x . El ancho de este intervalo de predicción toma en cuenta la incertidumbre asociada tanto con $\hat{\mu}$ como con el resultado de la distribución de respuesta.

2.2.1.7 Evaluación de ajustes y Devianza

La bondad del ajuste de un modelo a los datos es una pregunta natural que surge con todos los modelos estadísticos. Los principios de las pruebas de significancia, la selección de modelos y las pruebas de diagnóstico, analizados en los modelos de regresión Normal, son los mismos para los GLM. Sin embargo, algunos detalles técnicos de los métodos difieren en algo.

Una forma de evaluar el ajuste de un modelo dado es compararlo con el modelo con el mejor ajuste posible. El mejor ajuste se obtendrá cuando haya tantos parámetros como observaciones: a esto se le llama modelo saturado. Un modelo saturado asegurará que haya flexibilidad completa en el ajuste de θ_i . Desde que

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - \mu_i}{\phi} = \frac{y_i - \dot{a}(\theta_i)}{\phi},$$

el MLE de θ_i bajo el modelo saturado es $\check{\theta}_i$, donde $\dot{a}(\check{\theta}_i) = y_i$. Por lo tanto, el valor ajustado es igual a la observación y el modelo saturado se ajusta perfectamente. El valor del *log-likelihood* saturado es

$$\check{l} \equiv \sum_{i=1}^n \left[\ln c(y_i, \phi) + \frac{y_i \check{\theta}_i - a(\check{\theta}_i)}{\phi} \right],$$

el cual es el posible máximo *log-likelihood* para y dado la distribución de respuesta especificada por $a(\theta)$. Este valor es comparado a \hat{l} , el valor máximo del *log-likelihood* basado en y y las variables explicativas dadas. La Devianza denotada como Δ , es definido como una medida de distancia entre el modelo saturado y el modelo ajustado:

$$\Delta \equiv (\check{l} - \hat{l}).$$

- Cuando el modelo proporciona un buen ajuste, entonces se espera que \hat{l} sea cercano a (pero no mayor que) \check{l} . Un valor grande de la Devianza indica un modelo mal ajustado.
- El tamaño de Δ se evalúa en relación con la distribución χ^2_{n-p} . Esta es la distribu-

ción de muestreo aproximada de la desviación, suponiendo que el modelo ajustado es correcto y n es grande. El valor esperado de la Devianza es $n - p$. La regla de decisión para concluir si el modelo está mal ajustado es dividir Δ por sus grados de libertad $n - p$, y ver si se obtiene un valor mucho mayor que uno, de acuerdo al nivel crítico dado por el nivel significancia utilizado.

- Un cálculo directo muestra que para la familia exponencial

$$\Delta = 2 \sum_{i=1}^n \left[\frac{y_i(\check{\theta}_i - \hat{\theta}_i) - a(\check{\theta}_i) + a(\hat{\theta}_i)}{\phi} \right] \quad (2.10)$$

ya que los términos que involucran $c(y_i, \phi)$ se cancelan, y donde $\check{\theta}_i$ y $\hat{\theta}_i$ son tales que $\dot{a}(\check{\theta}_i) = y_i$ y $g[\dot{a}(\hat{\theta}_i)] = x'_i \hat{\beta}$, respectivamente.

- Cuando ϕ es desconocido y reemplazado por su estimador, entonces la distribución para la Devianza se ve comprometida. Para el caso de la distribución de Poisson tenemos que $\phi = 1$, y la aproximación χ^2 es útil. En el caso de la distribución normal si se conoce σ^2 , entonces la distribución χ^2 de la Devianza es exacta. Sin embargo, cuando se utiliza el estimador de σ^2 , no podemos confiar en que la Devianza se distribuya como χ^2 . *Hay que tener precaución en el uso de la Devianza como una medida general de bondad de ajuste*, puesto que su distribución aproximada a la χ^2_{n-p} depende de supuestos que con frecuencia no son sostenibles o son difíciles de verificar. Sin embargo, la Devianza es útil para probar la importancia de las variables explicativas en modelos anidados (Anderson et al., 2007; Garrido et al., 2016; Rao et al., 2010).

TABLA N° 2.3: Devianza para distribuciones de la familia exponencial

Distribución	Devianza Δ
Normal	$\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2$
Poisson	$2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right]$
Binomial	$2 \sum_{i=1}^n n_i \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) + (n_i - y_i) \ln \left(\frac{n_i - y_i}{1 - \hat{\mu}_i} \right) \right]$
Gamma	$2\nu \sum_{i=1}^n - \left[\ln \left(\frac{y_i}{\hat{\mu}_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$
Inversa Gausiana	$\frac{1}{\sigma^2} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2 y_i}$
Binomial Negativa	$2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i + \frac{1}{\kappa}) \ln \left(\frac{y_i + 1/\kappa}{\hat{\mu}_i + 1/\kappa} \right) \right]$

2.2.1.8 Prueba de significancia de las variables explicativas

Probar la significancia de las variables explicativas de modelos GLM es similar que en el Modelo Lineal General con residuos normales. Las hipótesis se escriben como $C\beta = r$, donde C es la matriz de hipótesis y r es un conjunto de valores dados.

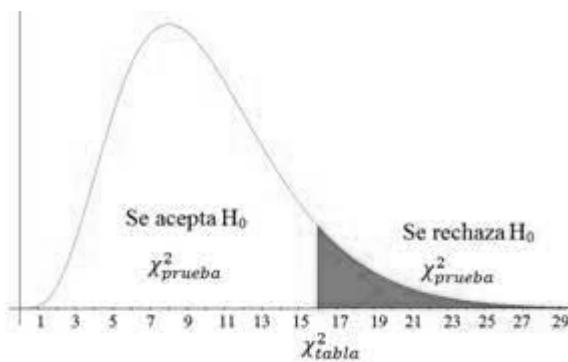
Existen diversos enfoques para probar hipótesis de la forma $C\beta = r$ y ver la significancia de las variables explicativas, pero todos ellos utilizan la *log-verosimilitud*, definido en la Sección 2.2.1.2. Denotamos por $\hat{\beta}$ el MLE no restringido de β , y con $\tilde{\beta}$ el MLE de β cuando se maximiza sujeto a las restricciones $C\beta = r$. Además, escribimos \hat{l} , como el valor de l , en $\hat{\beta}$, y \tilde{l} , el valor de l en $\tilde{\beta}$. Obviamente $\hat{l} \geq \tilde{l}$. A continuación describimos los métodos más utilizados para ver significancia de las variables.

Prueba del ratio de verosimilitud: Aquí \hat{l} es comparado con \tilde{l} . Un valor de \hat{l} mucho más grande que \tilde{l} es evidencia contra las restricciones. Ambos $\hat{\beta}$ y $\tilde{\beta}$ son requeridos. El ratio de verosimilitud es definido como $\lambda = \hat{L}/\tilde{L}$ donde \hat{L} y \tilde{L} son las verosimilitudes de los modelos restringidos y no restringidos, respectivamente. El estadístico de prueba de razón de verosimilitud es

$$2\ln\lambda = 2(\hat{l} - \tilde{l}) \quad (2.11)$$

Esto es siempre no-negativo, y tiene distribución χ_q^2 si $C\beta = r$, donde q es el número de filas de C , es decir, el número de restricciones sobre β . Si $2\ln\lambda$ es pequeño (cercano a cero) entonces el modelo restringido es casi tan bueno como el modelo no restringido, el cual podría proporcionar que $C\beta = r$. La región de rechazo para la prueba es la cola superior de la distribución χ_q^2 . La distribución χ_q^2 del estadístico de razón de verosimilitud involucra el

FIGURA N° 2.2: Distribución Chi-Cuadrado



Fuente: Elaboración Propia

parámetro de dispersión ϕ , que a menudo se desconoce. En el caso de las distribuciones de Poisson y Binomial ϕ es conocido; en el caso de otras distribuciones, generalmente se debe estimar ϕ . La distribución χ_q^2 sigue siendo apropiada en este caso, siempre que se utilice un estimador consistente para ϕ . El estadístico del ratio de verosimilitud puede ser expresado como la diferencia en las Devianzas de los modelos restringidos y no restringidos, siempre

que se use la misma estimación para ϕ en ambas *log – likelihoods*. Tenga en cuenta que la aproximación de la distribución de la Devianza por una de χ^2 puede ser cuestionable, excepto en el caso de distribución de Poisson donde se conoce ϕ (Rao et al., 2010).

Prueba de Wald: Esta prueba mide qué tan lejos está $C\hat{\beta}$ de r . Si la diferencia $C\hat{\beta} - r$ es grande, esto proporciona evidencia contra las restricciones. Sobre la hipótesis nula $H_0 : C\beta = r$, entonces como $\hat{\beta} = N[\beta, \phi(X'WX)^{-1}]$ esto sigue que

$$C\hat{\beta} - r \approx N[0, \phi C(X'WX)^{-1}C'].$$

Esto nos lleva al famoso estadístico de Wald para probar $C\beta = r$:

$$(C\hat{\beta} - r)' [\phi C(X'WX)^{-1}C']^{-1} (C\hat{\beta} - r) \approx \chi_q^2 \quad (2.12)$$

Por lo tanto, el estadístico de Wald es el cuadrado de la distancia estadística de $C\hat{\beta}$ de r utilizando la matriz de covarianzas de $C\hat{\beta}$. En la práctica, W se reemplaza por una estimación y por lo tanto, la distribución de χ_q^2 es aproximada (Rao et al., 2010).

2.2.2 Cúpulas

2.2.2.1 Definición

Actualmente, las Cúpulas son una herramienta fundamental que permiten escribir la función de distribución conjunta de un vector aleatorio como la composición de una cúpula y sus distribuciones marginales. En ese sentido, las cúpulas se han convertido en una herramienta importante para modelar vectores aleatorios donde las marginales son dependientes.

Definición 1. Una cúpula es una función de distribución multivariada $C : [0, 1]^n \rightarrow [0, 1]$, cuyas distribuciones marginales tienen distribución uniforme $U_i \sim U(0, 1), i = 1, \dots, n$. Las cúpulas satisfacen las siguientes propiedades:

1. $C(u_1, \dots, u_d)$ es creciente en cada componente u_i .
2. $C(u_1, \dots, 0, \dots, u_d) = 0$ para todo $u_i \in [0, 1]$.
3. $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ para todo $i \in \{1, \dots, d\}$, $u_i \in [0, 1]$.
4. Para todo $(a_1, \dots, a_d), (b_1, \dots, b_d) \in [0, 1]^d$ con $a_i \leq b_i$ tenemos que

$$\sum_{i_1=1}^2 \cdots \sum_{i_d=1}^2 (-1)^{i_1+\dots+i_d} C(u_{1i_1}, \dots, u_{di_d}) \geq 0$$

donde $u_{j1} = a_j$ y $u_{j2} = b_j$ para todo $j \in \{1, \dots, d\}$.

5. Para cada $u_i \in [0, 1]$, y para cada $j \in \{1, \dots, d\}$, la derivada parcial $\frac{\partial C}{\partial u_i}(u_1, \dots, u_d)$

existe es no negativa, es decir,

$$\frac{\partial C}{\partial u_i}(u_1, \dots, u_d) \geq 0.$$

Si F_i es la distribución acumulada de una variable aleatoria X_i , entonces

$$C(F_1(x_1), \dots, F_d(x_d))$$

es una distribución d -dimensional para el vector aleatorio $X = (X_1, \dots, X_d)$, con distribución marginal F_i , $i = 1, \dots, d$. De igual forma, si H es una función de distribución acumulada d -dimensional con marginales F_1, \dots, F_d , entonces existe una cópula d -dimensional C tal que, para todo $x = (x_1, \dots, x_d)$ se tiene

$$H(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)) \quad (2.13)$$

Si F_1, \dots, F_d son continuas, entonces C es único ([Sklar, 1959](#)), en otro caso, existen varias posibles cópulas, pero todas ellas coinciden sobre la clausura de $\text{Ran}(F_1) \times \dots \times \text{Ran}(F_d)$, donde $\text{Ran}(F)$ denota el rango de F (ver [Genest y Neslehova \(2007\)](#) para una demostración).

Teorema 1. Sea C una cópula. Entonces para todo $(u_1, \dots, u_n), (v_1, \dots, v_n) \in [0, 1]^n$ se tiene que:

$$|C(v_1, \dots, v_n) - C(u_1, \dots, u_n)| \leq \sum_{k=1}^n |v_k - u_k|.$$

La desigualdad anterior muestra que toda cópula C es uniformemente continua sobre $[0, 1]^n$.

Una demostración del Teorema 1 puede ser encontrada en ([Nelsen, 2006](#)).

El siguiente teorema nos muestra otra propiedad importante de las cópulas.

Teorema 2. Dada una cópula C , para todo $(u_1, \dots, u_n) \in [0, 1]^n$, la derivada parcial $\partial C / \partial \mu_i$ existe para casi todo $u_i \in [0, 1]$ con $i \in 1, \dots, n$. Además tenemos que

$$0 \leq \frac{\partial}{\partial u_i} C(u_1, \dots, u_n) \leq 1.$$

La demostración de este teorema puede ser encontrada en ([Nelsen, 2006](#)).

2.2.2.2 Teorema de Sklar

En un importante trabajo ([Sklar, 1959](#)), Sklar demostró que es posible describir de forma completa la estructura de dependencia que existe en un vector aleatorio a través de las distribuciones marginales y una función link que fue denominada *cópula*. Este resultado demostró

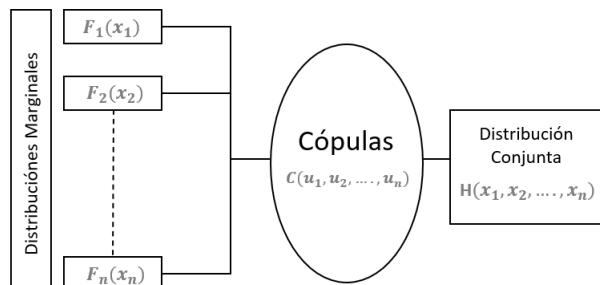
que toda distribución multivariada se puede generar a partir de una cópula y las distribuciones marginales, siendo una herramienta importante para diversas aplicaciones reales, donde se conoce las distribuciones marginales pero no se tiene información sobre la estructura de dependencia entre ellas. A continuación presentamos el teorema principal de la teoría de cópulas.

Teorema 3 (Teorema de SKLAR, Sklar (1959)). Sea H la distribución conjunta de las variables aleatorias X_1, \dots, X_n con marginales F_1, \dots, F_n . Entonces existe una cópula C tal que

$$H(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)), \quad \text{para todo } x_1, \dots, x_n \in \overline{\mathbb{R}}^1$$

Si F_1, \dots, F_n son continuas, entonces la cópula C es única. De otra forma, C está definida solamente en el $\text{Rango}F_1 \times \dots \times \text{Rango}F_n$. Por otro lado, si C es una cópula y F_1, \dots, F_n son funciones de distribución, entonces la función H es una función de distribución conjunta con marginales F_1, \dots, F_n .

FIGURA N° 2.3: Descripción de como se utiliza el Teorema de Sklar para calcular la distribución conjunta de un vector aleatorio.



Fuente: Elaboración propia.

En la Figura 2.16, mostramos gráficamente una distribución bivariada. En las diversas aplicaciones a datos reales lo que generalmente se observa son las distribuciones marginales. Partiendo de la información muestral de las marginales se quiere construir la distribución conjunta. Las diversas familias de cópulas que existen, ayudan a encontrar una aproximación de la verdadera distribución conjunta.

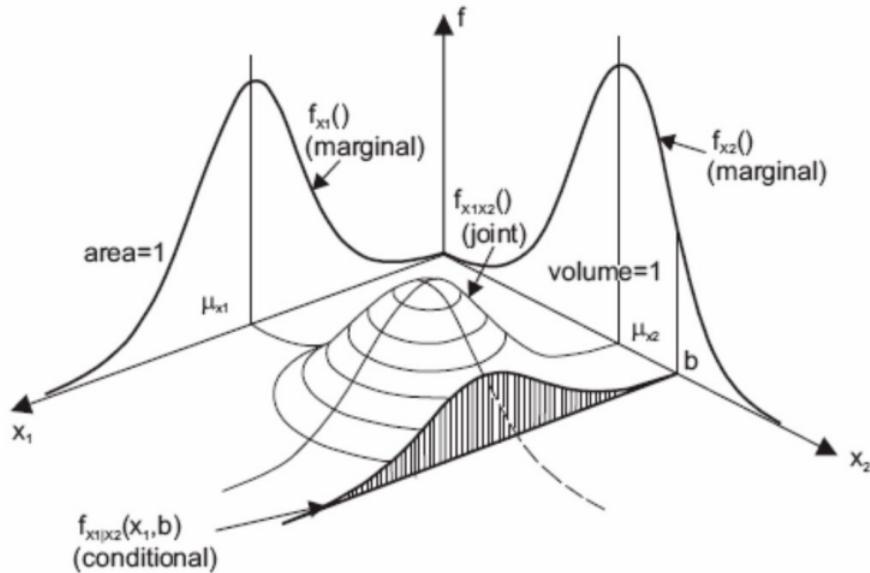
2.2.2.3 Cópulas y variables aleatorias

A continuación damos algunas definiciones necesarias para el desarrollo de este trabajo.

Definición 2. Dado un espacio de probabilidad $(\Omega, \mathfrak{F}, P)$, decimos que la función $X : (\Omega, \mathfrak{F}) \rightarrow (\mathbb{R}, \mathfrak{B})$ es una variable aleatoria, si X es una función medible, donde \mathfrak{B} es el σ -álgebra de Borel.

¹ $\overline{\mathbb{R}}$ denota la recta real $[-\infty, \infty]$

FIGURA N° 2.4: Función de densidad conjunta



Fuente: Figura tomada de Uncertainty treatment in civil engineering numerical models ([Matos et al., 2007](#))

Proposición: Decimos que las variables aleatorias X_1, \dots, X_n son independientes si, y sólo si, el producto de sus funciones de distribución $F_1(x_1), \dots, F_n(x_n)$ es igual a su función de distribución acumulada conjunta $H(x_1, \dots, x_n)$,

$$H(x_1, \dots, x_n) = F_1(x_1) \times \dots \times F_n(x_n), \quad \text{para todo } x_1, \dots, x_n \in \bar{\mathbb{R}}.$$

Teorema 4. Sean las variables aleatorias X_1, \dots, X_n con funciones de distribución continuas dadas por F_1, \dots, F_n , y función de distribución conjunta H . Por Teorema de Sklar existe una única cópula C tal que

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = H(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)).$$

Entonces, X_1, \dots, X_n son independientes si, y sólo si,

$$C = \prod_{i=1}^n u_i = u_1 \times \dots \times u_n$$

Para una demostración del Teorema 4, ver por ejemplo ([Nelsen, 2006](#)).

Teorema 5. Sea $(X_1, \dots, X_n)^T$ un vector aleatorio cuyas coordenadas son continuas, y sea C su respectiva cópula. Si $\alpha_1, \dots, \alpha_n$ son funciones reales estrictamente crecientes con dominios dados por $Rango X_1, \dots, Rango X_n$, respectivamente. Entonces, el vector aleatorio $(\alpha_1(X_1), \dots, \alpha_n(X_n))$ tiene cópula C . Es decir, C es invariante bajo transformaciones es-

trictamente crecientes de X_1, \dots, X_n .

Para una demostración del Teorema 5, ver por ejemplo ([Nelsen, 2006](#)).

Esta propiedad será muy útil en la aplicación que veremos más adelante.

2.2.2.4 Medidas de Dependencia

Las relaciones de dependencia entre variables aleatorias es uno de los temas más estudiados en probabilidad y estadística. La naturaleza de la dependencia puede tomar una variedad de formas y, a menos que se hagan algunos supuestos específicos sobre la dependencia, no se puede contemplar ningún modelo estadístico significativo. Por ese motivo, modelar la dependencia entre variables aleatorias es un problema complicado de abordar actualmente, desde el punto de vista estadístico y numérico.

Concordancia: Informalmente, un par de variables aleatorias son concordantes si los valores grandes de una variables tiende a estar asociado a valores grandes de la otra, y valores pequeños de uno con los valores pequeños del otro. Para ser más preciso, sean (x_i, y_i) y (x_j, y_j) dos observaciones de un vector aleatorio continuo (X, Y) . Decimos que (x_i, y_i) y (x_j, y_j) son concordantes si $x_i < x_j$ y $y_i < y_j$, o si $x_i > x_j$ y $y_i > y_j$. Similarmente, decimos que (x_i, y_i) y (x_j, y_j) son discordantes si $x_i < x_j$ y $y_i > y_j$ o $x_i > x_j$ y $y_i < y_j$. Una formulación alternativa es dada por: (x_i, y_i) y (x_j, y_j) son concordantes si $(x_i - x_j)(y_i - y_j) > 0$ y son discordantes si $(x_i - x_j)(y_i - y_j) < 0$. (*ver Nelsen (2006) para más detalles y propiedades de esta definición*)

Teorema 6. Sea (X, Y) y (\tilde{X}, \tilde{Y}) dos vectores aleatorios independientes continuas con función de distribución conjunta H y \tilde{H} , respectivamente, y con marginales comunes F (de X y \tilde{X}) G (de Y y \tilde{Y}). Sea C y \tilde{C} las cópulas de (X, Y) y (\tilde{X}, \tilde{Y}) respectivamente. Así tenemos que $H(x, y) = C(F(x), G(y))$ y $\tilde{H} = \tilde{C}(F(x), G(y))$. Sea Q la diferencia entre la probabilidad de concordancia y discordancia de (X, Y) y (\tilde{X}, \tilde{Y}) , es decir,

$$Q = P[(X - \tilde{X})(Y - \tilde{Y}) > 0] - P[(X - \tilde{X})(Y - \tilde{Y}) < 0]$$

Entonces,

$$Q = Q(C, \tilde{C}) = 4 \iint_{[0,1]^2} \tilde{C}(u, v) dC(u, v) - 1$$

Para una demostración del Teorema 6, ver ([Nelsen, 2006](#)).

Definición 3 ([Scarsini \(1984\)](#)). Una medida de dependencia entre dos variables aleatorias continuas X e Y , denotada por κ , cuya cópula es C , es una medida de concordancia tal que:

1. κ está definida para todo par de variables aleatorias X, Y .
2. $-1 \leq \kappa_{X,Y} \leq 1$, $\kappa_{X,X} = 1$ y $\kappa_{X,-X} = -1$

3. $\kappa_{X,Y} = \kappa_{Y,X}$
4. Si X e Y son independientes, entonces $\kappa_{X,Y} = 0$.
5. $\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}$
6. Si C y \tilde{C} son cópulas tales que $C \prec \tilde{C}$, entonces $\kappa_C \leq \kappa_{\tilde{C}}$
7. Si (X_n, Y_n) es una sucesión de vectores aleatorios continuos con cópulas C_n . Si C_n converge a C , tenemos que $\lim_{n \rightarrow \infty} \kappa_{C_n} = \kappa_C$

De la Definición 3 tenemos que si Y es una función no decreciente de X , entonces $\kappa_{X,Y} = \kappa_M = 1$ (*Comonotonicidad*), y si Y es una función monótona decreciente de X , entonces $\kappa_{X,Y} = \kappa_W = -1$ (*Contramonitoronoticidad*). Si α y β son funciones estrictamente crecientes en el *Rango(X)* y *Rango(Y)*, respectivamente, entonces $\kappa_{\alpha(X),\beta(Y)} = \kappa_{X,Y}$.

Correlación de Pearson ρ : Es una de las medidas de dependencia más utilizadas, el cual mide la relación lineal que existe entre variables aleatorias. Para dos variables aleatorias X e Y la correlación de Pearson $\rho(X, Y)$, se define por:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

Por su fácil implementación, el coeficiente de correlación de Pearson es aceptado como una medida de dependencia de amplia aceptación. Su practicidad se debe a que sólo se necesita estimar los dos primeros momentos de los datos.

Otra ventaja que tiene la correlación de Pearson es su relación con la función de distribución normal multivariada, donde la correlación resume toda la relación de dependencia existente. Sin embargo, se debe tener cuidado cuando se utiliza la correlación de Pearson como medida de dependencia, puesto que no es adecuada para medir dependencia no lineal (*ver Embrechts (1999), McNeil et al. (2005)*). Otros inconvenientes de la correlación de Pearson son:

- Si las variables son normales y el coeficiente de correlación de Pearson es cero, entonces las variables son independientes. Para otras distribuciones esta propiedad no necesariamente es verdad.
- Únicamente se define para dos de variables aleatorias que tengan varianzas finitas. Esto es de vital importancia, cuando se analiza variables con colas pesadas, puesto que en estos casos la varianza puede no existir.
- La correlación de Pearson no es invariantes bajo transformaciones estrictamente monótonas.

Una forma gráfica de entender que el coeficiente de correlación no es óptimo para medir la relación o asociación entre dos variables es mediante el cuarteto de Anscombe ([Anscombe](#),

1973). Por ejemplo, consideremos los siguientes datos de la [Tabla N° 2.4](#). Podemos observar que las variables X tienen la misma media al igual que las variables Y . Además, podemos observar que los coeficientes de correlación de cada grupo es el mismo.

TABLA N° 2.4: Datos del Cuarteto de Anscombe

	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
	X	Y	X	Y	X	Y	X	Y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58	
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76	
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71	
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84	
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47	
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04	
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25	
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50	
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56	
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91	
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89	
Media	9.00	11.00	9.00	11.00	9.00	11.00	9.00	11.00
ρ	0.816		0.816		0.816		0.816	

En la [Figura N° 2.5](#) mostramos la gráfica de dispersión de los 4 grupos de datos mostrados en la [Tabla N° 2.4](#). El primer gráfico (arriba a la izquierda) muestra una típica relación lineal simple entre dos variables correlacionadas cumpliendo con la suposición de normalidad. En el segundo gráfico (arriba a la derecha) se observa que los datos no están relacionados de forma lineal, y el coeficiente de correlación de Pearson (que mide relación lineal) no sería apropiado. En la tercera gráfica (abajo a la izquierda) la distribución de los datos es lineal, pero con una recta de regresión diferente a las anteriores por causa del dato extremo observado, que influye lo suficiente como para alterar la recta de regresión y disminuir el coeficiente de correlación de 1 a 0.816. Por último, la cuarta gráfica (abajo a la derecha) es un ejemplo donde un valor atípico es suficiente para producir un coeficiente de correlación alto incluso cuando la relación entre las dos variables no es lineal.

El cuarteto muestra la importancia de visualizar gráficamente los conjunto datos antes de ajustar algún tipo de modelo.

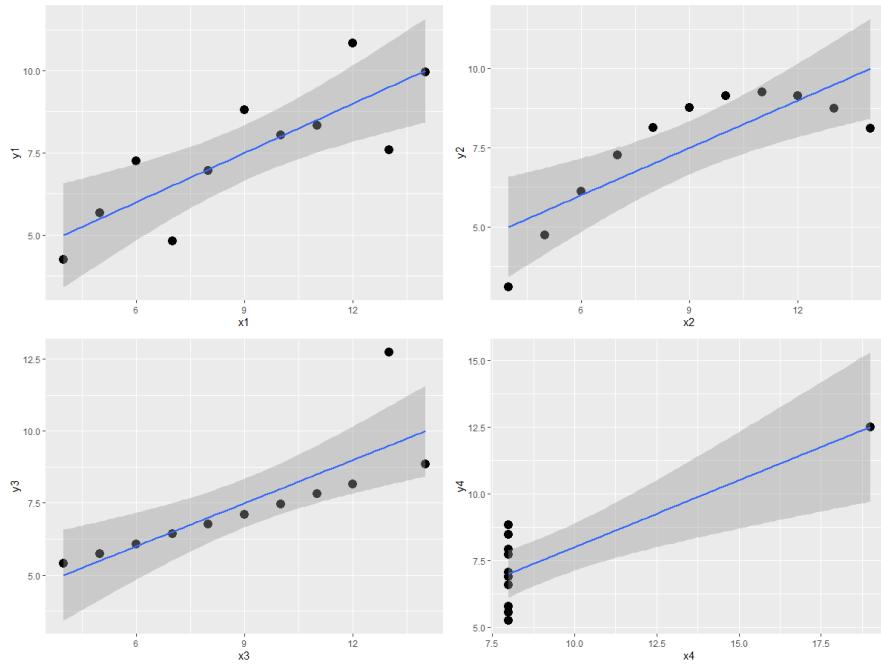
La Tau de Kendall τ : Sea $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ una muestra aleatoria de n observaciones de un vector aleatorio continuo (X, Y) . Un par de observaciones (x_i, y_i) y (x_j, y_j) son concordantes o discordantes según la definición de Concordancia, dada anteriormente.

Existen $\binom{n}{2}$ diferentes pares (x_i, y_i) y (x_j, y_j) de observaciones, y cada par es concordante o discordante. Sea N_c el número de pares concordantes y N_d el número de pares discordantes. Entonces la estimación del τ de Kendall para la muestra se define como:

$$\hat{\tau} = \frac{N_c - N_d}{N_c + N_d}$$

Equivalentemente el τ de Kendall se define como la probabilidad de concordancia menos

FIGURA N° 2.5: Gráficos del Cuarteto de Anscombe



la probabilidad de discordancia para un par de observaciones (x_i, y_i) y (x_j, y_j) elegidas aleatoriamente de la muestra.

Siguiendo la definición anterior damos la definición del tau de Kendall para un vector aleatorio.

Definición 4. Definimos el tau de Kendall de vector aleatorio (X, Y) de la siguiente forma:

$$\tau(X, Y) = P \left[(X - \tilde{X})(Y - \tilde{Y}) > 0 \right] - P \left[(X - \tilde{X})(Y - \tilde{Y}) < 0 \right],$$

donde (\tilde{X}, \tilde{Y}) es una copia independiente de (X, Y) .

El siguiente teorema da una forma equivalente de calcular la tau de Kendall en función de la cópula de X e Y .

Teorema 7. Sea un vector aleatorio continuo (X, Y) con cópula C . Entonces la tau de Kendall es definido por

$$\tau_C = 4 \iint_{[0,1]^2} C(u, v) dC(u, v) - 1.$$

Observe que la tau de Kendall se puede escribir en función del valor esperado de la variable aleatoria $C(U, V)$, donde $U, V \sim U(0, 1)$, es decir,

$$\tau_C = 4E[C(U, V)] - 1$$

2.2.2.5 Medidas gráficas para detectar dependencia bivariante

Chi-Plot: En Fisher y Switzer (2001), los autores construyen una medida gráfica como instrumento sencillo, rápido y útil para explorar estructuras de dependencia entre dos variables, dependiendo únicamente de los datos. La información que proporciona esta medida complementa a la obtenida de un diagrama de dispersión, en el cual puede resultar difícil de observar algún patrón de dependencia. Para su construcción procedemos de la siguiente manera: Sea $(X_1, Y_1), \dots, (X_n, Y_n)$ una muestra aleatoria simple de H , función de distribución conjunta continua de la v.a. (X, Y) . Denotamos por $I(A)$ la función indicadora del evento A .

Para cada observación x_i, y_i , seguimos los siguientes pasos:

- Calculamos H_i, F_i, G_i y S_i que son definidos por:

$$H_i = \frac{1}{n-1} \sum_{j \neq i} I(X_j \leq X_i, Y_j \leq Y_i)$$

$$F_i = \frac{1}{n-1} \sum_{j \neq i} I(X_j \leq X_i)$$

$$G_i = \frac{1}{n-1} \sum_{j \neq i} I(Y_j \leq Y_i)$$

$$S_i = \text{sign} \left\{ \left(F_i - \frac{1}{2} \right) \left(G_i - \frac{1}{2} \right) \right\}$$

- Calculamos λ_i y χ_i de la siguiente forma:

$$\lambda_i = 4S_i \max \left\{ \left(F_i - \frac{1}{2} \right)^2, \left(G_i - \frac{1}{2} \right)^2 \right\}$$

$$\chi_i = \frac{H_i - F_i G_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}}$$

El Chi-plot es un diagrama de los pares $(\lambda_i, \chi_i)_{i=1, \dots, n}$ donde λ_i es una medida de distancia de la observación (X_i, Y_i) al centro de los datos. Los valores de λ_i deben estar en el intervalo $[-1, 1]$.

En caso de que los datos constituyan una muestra bivariante con marginales continuas independientes, los valores de λ_i se distribuyen uniformemente. Sin embargo, si X e Y son dependientes, los valores de λ_i se presentarán formando de clusters o agrupaciones. En particular, valores positivos de λ_i indican que X_i e Y_i son relativamente grandes (a la vez) o relativamente pequeñas (a la vez) respecto a sus medianas, mientras que λ_i negativos corresponden a X_i e Y_i situados en lados opuestos respecto a sus medianas.

Para cada muestra, χ_i se puede interpretar como el coeficiente de correlación asociado a $n - 1$ pares (X_{ij}, Y_{ij}) , definidos de la siguiente forma:

$$X_{ij} = \begin{cases} 1 & \text{Si } X_j \leq X_i \\ 0 & \text{en otro caso} \end{cases}$$

$$Y_{ij} = \begin{cases} 1 & \text{Si } Y_j \leq Y_i \\ 0 & \text{en otro caso} \end{cases}$$

para todo $j \neq i$. Así, tenemos que $-1 \leq \chi_i \leq 1$, para todo $i = 1, \dots, n$. Además, la expresión $\sqrt{n\chi_i}$ es la raíz cuadrada del estadístico chi-cuadrado utilizado tradicionalmente para contrastar la independencia en la tabla de contingencia generada por la muestra (X_i, Y_i) .

- Graficar los pares de puntos $(\lambda_i, \chi_i) \in [-1, 1] \times [-1, 1]$

De acuerdo a [Fisher y Switzer \(2001\)](#), para evitar posibles desfases gráficos, se recomienda graficar únicamente los puntos que satisfacen

$$|\lambda| \leq 4 \left(\frac{1}{n-1} - \frac{1}{2} \right)^2$$

Al momento de graficar la nube de puntos, se debe tener en cuenta la posible pérdida de pares de puntos (entre dos y cuatro, en general) al dividir por cero, por lo que este método no es eficiente en muestras pequeñas.

Si queremos contrastar las hipótesis $H_0: \text{Las variables son independientes}$ frente a $H_1: \text{Las variables no son independientes}$, los valores de χ_i que estén “muy lejos” de cero nos darán evidencia estadística para poder rechazar la hipótesis nula.

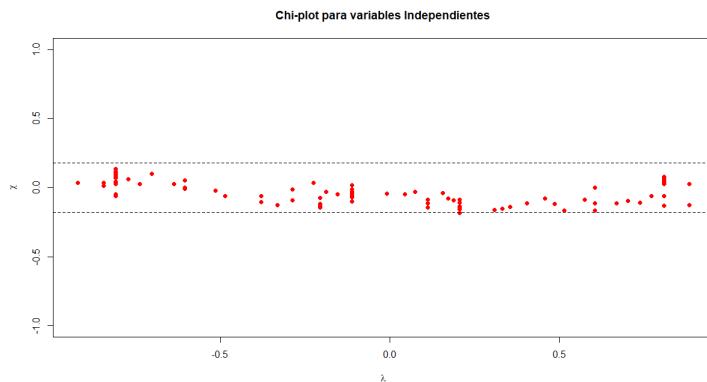
El término “muy lejos” está asociado con el nivel de significancia que se define a priori en la prueba de hipótesis. [Fisher y Switzer \(2001\)](#) proponen los siguientes límites de control $\pm c_p \sqrt{n}$, donde c_p es escogido aproximadamente como el $100p\%$ ² de los pares (λ_i, χ_i) que caen dentro de esas líneas.

Para ilustrar la técnica visual del estudio de la dependencia dada por la representación de Chi-plots, primero generamos una muestra aleatoria simple de tamaño 100 de dos variables independientes, una extraída de una variable $X \sim Poi(3)$ y otra $Y \sim N(0, 1)$, y calculamos a partir de ellas los (λ_i, χ_i) . De esta forma, se logran 100 pares de observaciones provenientes de dos variables independientes.

Ahora generamos una muestra aleatoria de dos variables que son dependientes, dando explícitamente una forma funcional a su dependencia. Para esto, escogemos $X \sim Poi(3)$ e

² p es equivalente al nivel de confianza. Se suele escoger $p = 0.9$ $p = 0.95$ $p = 0.99$

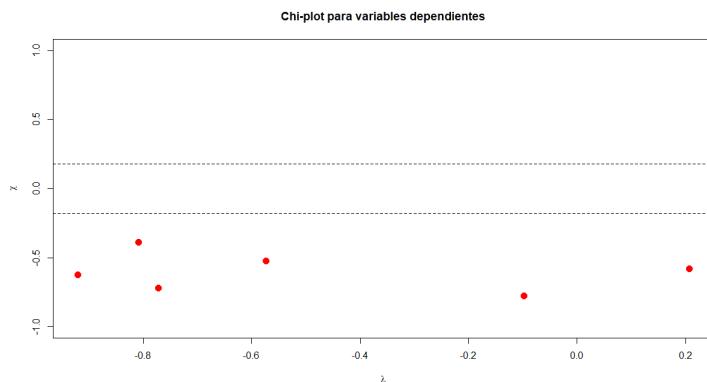
FIGURA N° 2.6: Chi-plot para variables independientes con $p = 0.95$



Fuente: Elaboración propia.

$Y = \frac{1}{1+X^2}$. Al tomar dos v.a. dependientes lo que obtendremos será una nube de puntos totalmente alejada de los márgenes de tolerancia dados por $\pm c_p \sqrt{n}$, puesto que las dos variables fueron escogidas dependientes. En la Figura 2.7 se puede observar claramente como la nube de puntos cae fuera de las bandas de tolerancia. Esto demuestraría que la muestra observada no es independiente, tal como fue tomada a priori.

FIGURA N° 2.7: Chi-plot para variables dependientes con $p = 0.95$



Fuente: Elaboración Propia

Para más detalles del método Chi-Plot, el lector puede revisar [Fisher y Switzer \(2001\)](#), que es el trabajo donde fue introducido este método, y de donde nos guiamos para desarrollar esta sección.

K-Plot: También conocida como Kendall-Plot. Esta es técnica gráfica que se utiliza para analizar la posible dependencia entre dos variables aleatorias fue propuesta por Genest y Boies en ([Genest y Boies, 2003](#)). Este método está inspirado en los famosos gráficos QQ-plots. A continuación pasamos a describir el método:

Primero representamos los pares $(W_{i:n}, H_{(i)})$ para todo $i \in 1, \dots, n$ de modo que

$$H_{(1)} \leq H_{(2)} \leq \dots \leq H_{(n)}$$

es la muestra aleatoria ordenada, donde cada H_i es calculada como en el método Chi-plot.

Luego calculamos

$$W_{i:n} = n \cdot \binom{n-1}{i-1} \int_0^1 w (K_0(w))^{i-1} \cdot (1 - K_0(w))^{n-i} dK_0 w,$$

donde

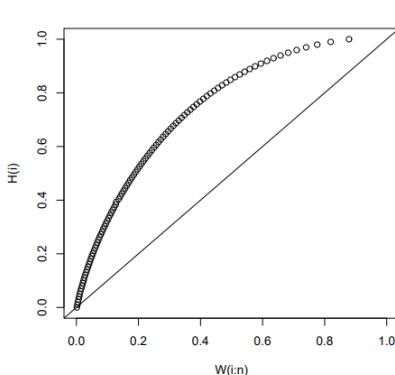
$$K_0(w) = P(U \cdot V \leq w) = w - w \cdot \log(w) \quad w \in [0, 1]$$

A medida que los pares $(W_{i:n}, H_{(i)})$ se vayan desviando con respecto a la diagonal, se puede ir asumiendo que existe una dependencia funcional entre las dos variables aleatorias. Cuando el desvío es mayor, eso implica que la dependencia también lo es.

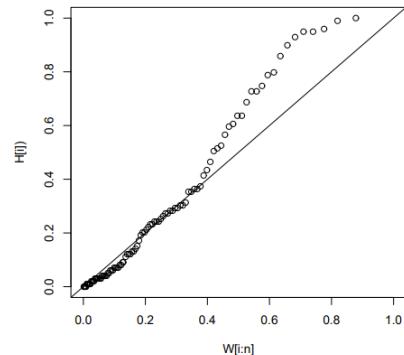
Para ilustrar este método consideremos una muestra de tamaño 100, de las variables X, Y , donde $Y = X^3$ y $X \sim N(0, 1)$. En la Figura 2.8a observamos que la nube de puntos está por encima de la diagonal. Sin embargo, cuando las variables resultan ser independientes la gráfica se concentra en la diagonal. Por ejemplo, en la Figura 2.8b ilustramos un K-plot para dos variables independientes $X \sim N(0, 1)$ e $Y \sim Po(3)$.

FIGURA N° 2.8: Gráficos K-PLOT

(a) Variables dependientes



(b) Variables independientes



2.2.3 Familias de Cúpulas

2.2.3.1 Cúpulas Elípticas

Su principal característica es que representan relaciones de dependencia simétricas sin importar si el análisis se realiza sobre la cola izquierda o derecha de las distribuciones implicadas. Las principales cúpulas elípticas son:

Cópula Normal o Gaussiana: Sea \mathbf{R} es la matriz de correlación de dimensión $d \times d$. Entonces la función de distribución acumulada de la cópula normal multivariante es:

$$C(u; \mathbf{R}) = \Phi_d(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d); \mathbf{R}), \quad u \in [0, 1]^d$$

y la densidad de la cópula es

$$c(u; \mathbf{R}) = \frac{\phi_d(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))}{\prod_{j=1}^d \phi(\Phi^{-1}(u_j))}, \quad u \in (0, 1)^d,$$

cuando \mathbf{R} es definida positiva.

Cuando $d = 2$, el parámetro de la función de la distribución acumulada, con $\rho \in [-1, 1]$, nos conlleva a:

$$C(u, v, \rho) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho), \quad 0 < u, v < 1.$$

La distribución condicional es:

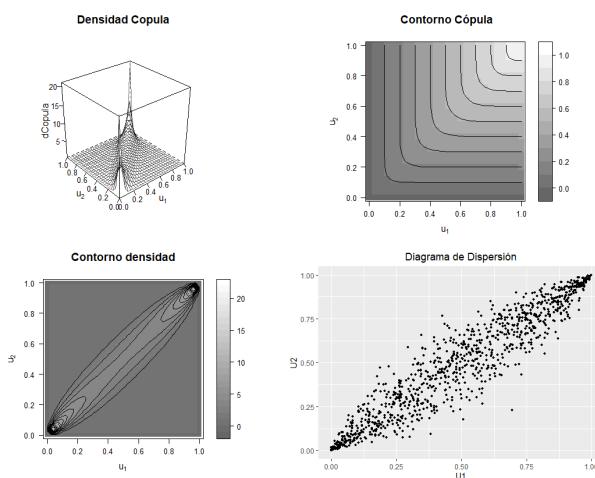
$$C_{2|1}(v|u; \rho) = \Phi\left(\frac{\Phi^{-1}(v) - \rho\Phi^{-1}(u)}{\sqrt{1-\rho^2}}\right).$$

y la densidad de la cópula, como es descrito en ([Joe, 2014](#)), con $-1 < \rho < 1$ es:

$$\begin{aligned} c(u, v; \rho) &= \frac{\phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho)}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))} \\ &= (1 - \rho^2)^{-1/2} \times \exp\left(-\frac{1}{2}(x^2 + y^2 - 2\rho xy)(1 - \rho^2)\right) \times \exp\left(\frac{1}{2}(x^2 + y^2)\right) \end{aligned}$$

Con $x = \Phi^{-1}(u)$, $y = \Phi^{-1}(v)$ y $0 < u, v < 1$.

FIGURA N° 2.9: Cópula Gaussiana ($\rho = 0.8$)



Elaboración propia usando el software R.

Cópula de Student: Es la cópula asociada a la distribución t multivariante. Para el caso bivariado tenemos que

$$C(u, v; \rho, \nu) = t_{\rho, \nu}(t^{-1}(u), t^{-1}(v))$$

Siendo t_ν la distribución de t -Student con ν grados de libertad y $t_{\rho, \nu}$ la distribución t -Student bivariada con ν grados de libertad y con matriz de correlación dada por ρ . La función de densidad de la cópula t bivariada se escribe como:

$$C(u, v; \rho, \nu) = \frac{\nu}{2\sqrt{1-\rho^2}} \cdot \frac{\Gamma(\frac{\nu}{2})^2}{\Gamma(\frac{\nu+1}{2})^2} \cdot \frac{\left(1 + \frac{x^2+y^2-2\rho xy}{\nu(1-\rho^2)}\right) - \frac{\nu+2}{2}}{\left[\left(1 + \frac{x^2}{\nu}\right)\left(1 + \frac{y^2}{\nu}\right)\right] - \frac{\nu+1}{2}}$$

Con $x = t_\nu^{-1}(u)$ y $y = t_\nu^{-1}(v)$. Para obtener lo anterior utilizamos la expresión de la distribución t bivariada:

$$t_{\rho, \nu}(x, y) = \int_{-\infty}^x \int_{-\infty}^y \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{x^2+y^2-2\rho xy}{\nu(1-\rho^2)}\right) - \frac{\nu+2}{2} dx dy$$

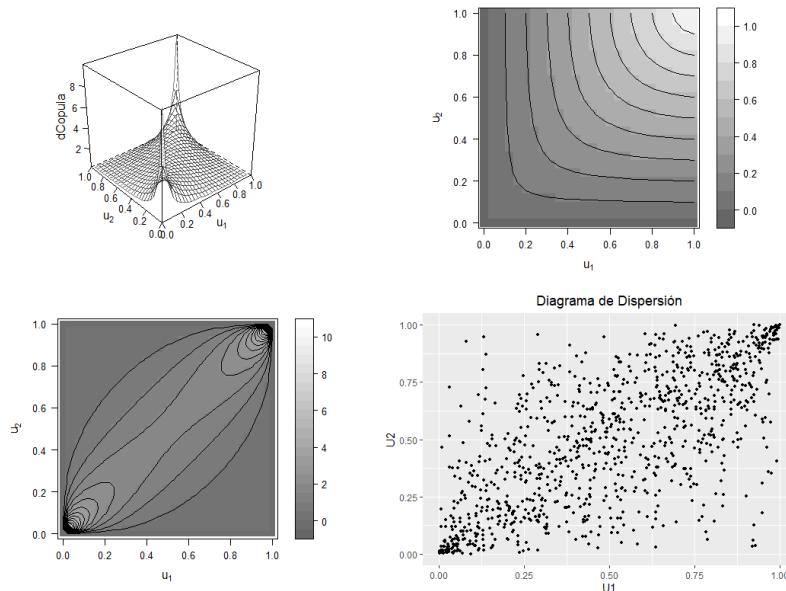
Luego utilizamos la siguiente propiedad de las cópulas

$$f(x, y) = C(F_1(x), F_2(y)) \cdot f_1(x) \cdot f_2(y)$$

con ello determinamos la expresión de la cópula t .

A menudo las cópulas elípticas son llamadas cópulas implícitas, puesto que no tienen una forma analítica, y solo se expresa en términos de las distribuciones bivariadas asociadas a los mismos.

FIGURA N° 2.10: Cópula Student($\tau = 0.5, \nu = 4$)



Elaboración propia usando el software R.

2.2.3.2 Cúpulas Arquimedias

Existe una gran diversidad de cúpulas que pertenecen a la familia arquimédiana y gracias a esta variedad permiten modelar muchos tipos de estructuras de dependencia. Otra ventaja de esta familia de cúpulas es que dependen solamente de un parámetro y son más fáciles de implementar numéricamente.

Definición 5. Sea $\varphi : [0, 1] \rightarrow [0, \infty]$ una función continua, estrictamente decreciente y convexa tal que $\varphi(0) = \infty$ y $\varphi(1) = 0$. Una cúpula arquimédiana es una cúpula que se puede representar de la siguiente forma:

$$C(u) = \varphi(\varphi^{-1}(u_1) + \dots + \varphi^{-1}(u_d)), \quad u \in [0, 1]^d,$$

La función φ recibe el nombre de **generador de la cúpula**.

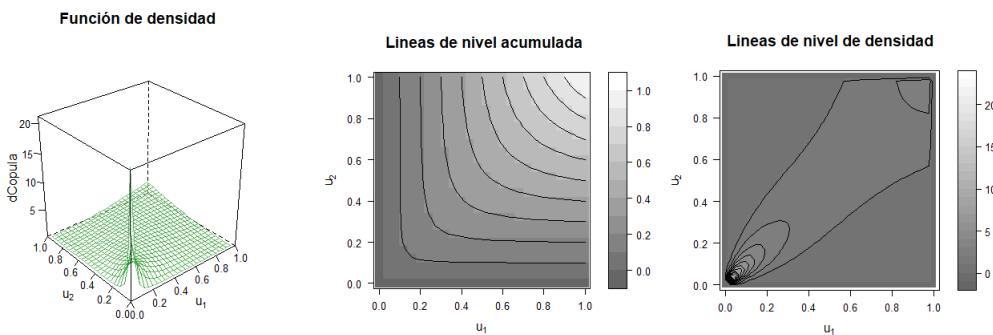
Cúpula de Clayton: Sea $\varphi_\theta(t) = \frac{t^{-\theta}-1}{\theta}$ para $\theta > 0$, el generador que corresponde a la familia de Clayton. La inversa $\varphi_\theta^{-1}(t) = (1+\theta t)^{-\frac{1}{\theta}}$ es estrictamente monótona sobre $[0, \infty]$, puesto que para todo $t \in [0, \infty)$, tenemos que

$$(-1)^k \frac{d^k}{dt^k} \varphi^{-1}(t) = (-1)^{2k} \prod_{j=0}^k (1+\theta j)(1+\theta t)^{-\frac{1+k\theta}{\theta}} \geq 0, \quad k = 0, 1, \dots$$

La cúpula Clayton multivariada, con $n \geq 2$, está dada por

$$C_\theta(u_1, \dots, u_n) := (u_1^{-\theta} + \dots + u_n^{-\theta} - n + 1)^{-\frac{1}{\theta}}, \quad \theta > 0.$$

FIGURA N° 2.11: Cúpula Clayton($\theta = 0.5$)



Elaboración propia usando el software R.

Cúpula de Frank: Sea $\varphi_\theta(t) = -\ln \frac{e^{-\theta t} - 1}{e^{-\theta} - 1}$, el generador de la familia de cúpulas de Frank. Aunque los generadores de esta familia son estrictos, debemos restringir los valores de θ a

$(0, \infty)$. La inversa del generador es igual a

$$\varphi_{\theta}^{-1}(t) = -\frac{1}{\theta} \ln[1 - (1 - e^{-\theta})e^{-t}].$$

Si probamos que $f(x) = -\frac{\ln(1-x)}{\theta}$ es absolutamente monótona en $(0, 1)$ y que $g(t) = (1 - e^{-\theta})e^{-t}$ es completamente monótona en $[0, \infty)$, $\varphi_{\theta}^{-1} = f \circ g$ será completamente monótona en $[0, \infty)$. Entonces,

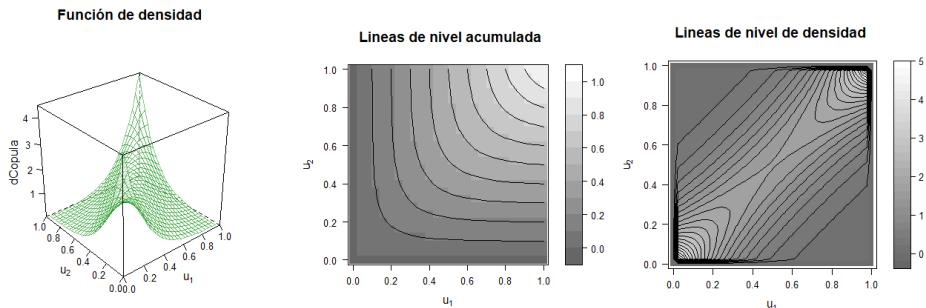
$$\frac{d^k}{dx^k} f(x) = \begin{cases} \frac{-\ln(1-x)}{\theta}, & k = 0 \\ \frac{(k-1)!}{\theta} (1-x)^{-k}, & k > 0 \end{cases}$$

$$(-1)^k \frac{d^k}{dt^k} g(t) = (-1)^{2k} (1 - e^{-\theta}) e^{-t} \geq 0.$$

Entonces, para $\theta > 0$ y $n \geq 2$, la cópula de Frank multivariada está dada por

$$C_{\theta}(u_1, \dots, u_n) := -\frac{1}{\theta} \ln \left(1 + \frac{(e^{-\theta u_1} - 1) \dots (e^{-\theta u_n} - 1)}{(e^{-\theta} - 1)^{n-1}} \right)$$

FIGURA N° 2.12: Cúpula de Frank($\theta = 0.5$)



Elaboración propia usando el software R.

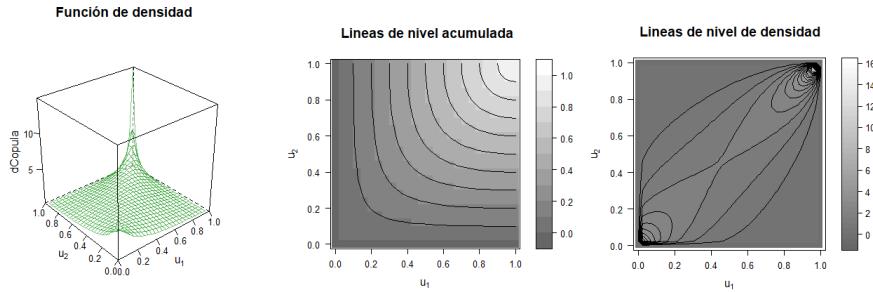
Cúpula de Gumbel: Sea $\varphi_{\theta} = (-\ln t)^{\theta}, \theta \geq 1$, el generador de la familia de Gumbel. Entonces $\varphi_{\theta}^{-1}(t) = \exp(-t^{\frac{1}{\theta}})$ es completamente monótona. En efecto, si g denota la derivada,

$$(-1)^k \frac{d^k}{dt^k} g(t) = (-1)^k \frac{\prod_{j=0}^k (1 - \theta j)}{\theta^{k+1}} t^{\frac{1-k\theta}{\theta}} \geq 0.$$

Entonces, para $\theta \geq 1$ y $n \geq 2$, la cópula de Gumbel multivariada está dada por

$$C_{\theta}(u_1, \dots, u_n) := \exp \left(- [(-\ln u_1)^{\theta} + \dots + (-\ln u_n)^{\theta}]^{\frac{1}{\theta}} \right)$$

FIGURA N° 2.13: Cúpula de Gumbel($\theta = 0.5$)



Elaboración propia usando el software R.

2.2.4 Modelos GLM basados en Cúpulas

2.2.4.1 Cúpulas bivariadas para datos continuos-discretos

Una cúpula bivariada $C : [0, 1] \times [0, 1] \rightarrow [0, 1]$ es una función de distribución acumulada bivariante en $[0, 1] \times [0, 1]$ con marginales distribuidas uniformemente. La importancia de la cúpula se sustenta mediante el Teorema de Sklar ([Sklar, 1959](#)). En el caso bivariado, establece que para cada función de distribución conjunta $F_{X,Y}$ de una variable aleatoria bivariada (X, Y) con funciones de distribución marginal F_X y F_Y , existe una cúpula bivariada C tal que

$$F_{X,Y}(x, y) = C(F_X(x), F_Y(y))$$

Si X e Y son variables aleatorias continuas, la cúpula C es única.

Si C es una cúpula, por el Teorema de Sklar $C(F_X(x), F_Y(y))$ define una función de distribución bivariada con marginales dadas por F_X y F_Y . Esto nos permite modelar las distribuciones marginales y la dependencia conjunta por separado, puesto que podemos definir la cúpula C independientemente de las distribuciones marginales.

Sabemos que las cúpulas son invariantes bajo transformaciones monótonas de las distribuciones marginales. Por lo tanto, en lugar del coeficiente de correlación de Pearson, que mide asociaciones lineales, se utilizan medidas de asociación monótonas. Una opción muy común es la τ de Kendall,

$$\tau := 4 \int_{[0,1]^2} C(u, v) dC(u, v) - 1 \in [-1, 1]$$

En ([Czado et al., 2012](#)) se estudian modelos basados en cúpula, donde se tiene una v.a. continua y una v.a. discreta. Sea X una variable aleatoria continua, e Y una variable aleatoria discreta. Su distribución conjunta es definida por una cúpula parámetrica $C(., .|\theta)$ que

depende del parámetro θ , es decir, la distribución conjunta está dada por

$$F_{X,Y|\theta}(x,y) = C(F_X(x), F_Y(y)|\theta)$$

Czado et al. (2012) usan principalmente cuatro familias de cópulas bivariadas parámetricas: Gauss, Clayton, Gumbel y Frank. Cada familia depende de un sólo parámetro θ . Este parámetro puede ser expresado en términos de la tau de Kendall como lo muestra la Tabla 2.5:

TABLA N° 2.5: Relación entre el parámetro θ y τ de Kendall

Familia	Cópula $C(u,v \theta)$	Rango de θ	Relación con τ
Gauss	$\Phi_2(\Phi^{-1}(u), \Phi^{-1}(v) \theta)$	$] -1, 1 [$	$\tau = \frac{2}{\pi} \arcsin(\theta) \in \mathbb{R}$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$] 0, \infty [$	$\tau = \frac{\theta}{\theta+2} \in] 0, \infty [$
Gumbel	$\exp\left(-((- \log u)^\theta + (- \log v)^\theta)^{1/\theta}\right)$	$[1, \infty [$	$\tau = \frac{\theta-1}{\theta} \in [0, \infty [$
Frank	$-\frac{1}{\theta} \log\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right)$	$\mathbb{R} - \{0\}$	$\tau = 1 - \frac{4}{\theta}[1 - D_1(\theta)] \in \mathbb{R} - \{0\}$

donde $D_k(x)$ es definido por

$$D_k(x) = \frac{k}{x^k} \int_0^x \frac{t^k}{e^t - 1} dt$$

denota la función Debye el cual está definida para $k \in \mathbb{N}$

En la Tabla 2.5 podemos observar que la cópula Clayton es sólo definida para valores positivos de la τ de Kendall, y que la cópula de Gumbel es sólo definida para valores no negativos de la τ de Kendall. Sin embargo, a través de una rotación, es posible extender estas familias de cópulas a valores negativos de τ .

Para el muestreo, la estimación y la predicción, necesitamos la función de densidad conjunta/masa de probabilidad de X e Y que se define por

$$f_{X,Y}(x,y) := \frac{\partial}{\partial x} P(X \leq x, Y = y) \quad (2.14)$$

Ahora derivamos las fórmulas para la densidad conjunta de X e Y en términos de la cópula $C(\cdot, \cdot | \theta)$, denotaremos lo siguiente:

$$C_1(u, v | \theta) := \frac{\partial}{\partial u} C(u, v | \theta) \quad (2.15)$$

para $u, v \in] 0, 1 [$ la derivada parcial de la cópula con respecto a la primera variable. Debemos notar que esto es la densidad condicional de la variable aleatoria $V := F_Y(Y)$ dado

$U := F_X(X)$. En la Tabla 2.6 mostramos las derivadas parciales para las cópulas Clayton, Gumbel, Frank y Gauss. Para mayor detalle de las derivadas parciales de las funciones cópulas revisar (Joe, 2014; Schepsmeier y Stöber, 2014).

Teorema 8 (Función de densidad). La función de densidad conjunta $f_{X,Y}$ de una variable aleatoria continua X y una variable aleatoria discreta Y puede escribirse en función de las marginales y la cópula C como,

$$f_{X,Y}(x,y|\theta) = f_X(x) \{C_1(F_X(x), F_Y(y) | \theta) - C_1(F_X(x), F_Y(y-1) | \theta)\}$$

(Para una demostración del Teorema 8, ver (Krämer et al., 2013).)

TABLA N° 2.6: Primera derivada parcial

Familia	$C_1(u,v \theta)$
Gauss	$\Phi\left(\frac{\Phi^{-1}(v) - \theta\Phi^{-1}(u)}{\sqrt{1-\theta^2}}\right)$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta-1} u^{-\theta-1}$
Gumbel	$u^{-1}\exp\left(-\left((-log u)^\theta + (-log v)^\theta\right)^{1/\theta}\right)$
Frank	$\frac{e^\theta(e^{\theta v}-1)}{e^{\theta(u+1)}+e^{\theta(v+1)}-e^\theta-e^{\theta(u+v)}}$

2.2.4.2 Distribuciones Marginales

Como una aplicación a la industria de seguros en Krämer et al. (2013) y Kholifah et al. (2019) los autores consideran el costo del siniestro X como variable aleatoria continua Gamma, y el número de siniestros Y como variable discreta. Para el desarrollo de este trabajo, utilizaremos las mismas distribuciones para la siniestralidad y para la frecuencia, propuestas en dichos trabajos. Es por ello que para el costo del siniestro utilizaremos el siguiente modelo:

$$f_X(x|\mu, \delta) = \frac{1}{x\Gamma\left(\frac{1}{\delta}\right)} \left(\frac{x}{\mu\delta}\right)^{\frac{1}{\delta}} \exp\left(-\frac{x}{\mu\delta}\right) \quad \text{para } x > 0$$

con media $\mu > 0$ y parámetro de dispersión dado por $\delta > 0$.

Para el número de siniestros Y , esta será modelada como una distribución Poisson con pa-

rámetro $\lambda > 0$,

$$f_Y(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \quad \text{para } y = 0, 1, 2, \dots$$

TABLA N° 2.7: Parámetros del modelo de distribución conjunta

	Severidad	Frecuencia	Familia de cópulas
Distribución	Gamma	Poisson	Gauss, Clayton, Gumbel, Frank
Parámetros	$\mu > 0, \delta > 0$	$\lambda > 0$	$\theta \in \Theta$
Esperanza	$E(X) = \mu$	$E(Y) = \lambda$	—
Varianza	$Var(X) = \mu^2 \delta$	$Var(Y) = \lambda$	—

2.2.4.3 Modelo de cópula conjunta para la frecuencia y severidad

Combinando las distribuciones marginales y el enfoque de cópula, obtenemos el siguiente modelo general.

Definición 6 (Modelo conjunto de la frecuencia y severidad). El modelo Gamma-Poisson basado en cópula mixta para la frecuencia Y y la severidad X se define mediante la función de densidad conjunta

$$f_{X,Y}(x,y|\mu,\delta,\lambda,\theta) = f_X(x|\mu,\delta) \{C_1(F_X(x|\mu,\delta), F_Y(y|\lambda)|\theta) - C_1(F_X(x|\mu,\delta), F_Y(y-1|\lambda)|\theta)\}$$

para $x > 0$ y $y = 0, 1, 2, \dots$

El modelo depende de cuatro parámetros: los parámetros μ, δ (Gamma) y λ (Poisson) para las distribuciones marginales, y el parámetro de la cópula θ . La Tabla 2.7 muestra los parámetros y sus relaciones con la distribución conjunta.

Teorema 9 (Distribución condicional). La distribución condicional de $Y|X=x$, es decir el número de siniestros dado un costo medio bajo el modelo basado en cópula de la Definicion 6, está dada por

$$P(Y=y|X=x, \mu, \delta, \lambda, \theta) = C_1(F_X(x|\mu, \delta), F_Y(y|\lambda)|\theta) - C_1(F_X(x|\mu, \delta), F_Y(y-1|\lambda)|\theta)$$

Para una demostración del Teorema 9, véase *Krämer et al. (2013)*.

2.2.4.4 Estimación de la pérdida póliza: Prima de riesgo conjunta

Definición 7 (Pérdida póliza). Para una póliza con costo medio X y número de siniestros Y , la pérdida póliza es definida como el producto de las cantidades mencionadas

$$L := X \cdot Y$$

La pérdida póliza es una variable aleatoria continua y positiva, y esta depende de los parámetros mostrados en la Tabla 2.7.

Basado en la definición previa, Krämer et al. (2013) demuestran el siguiente teorema, que describe la distribución de la pérdida póliza en función de la cópula y las distribuciones marginales.

Teorema 10 (función de densidad de la pérdida póliza). La distribución de la pérdida póliza L está dada por la siguiente función de densidad

$$f_L(l | \mu, \delta, \lambda, \theta) = \sum_{y=1}^{\infty} \left[C_1 \left(F_X \left(\frac{l}{y} | \mu, \delta \right), F_Y(y|\lambda) | \theta \right) - C_1 \left(F_X \left(\frac{l}{y} | \mu, \delta \right), F_Y(y-1|\lambda) | \theta \right) \right] \cdot \frac{1}{y} f_X \left(\frac{l}{y} | \mu, \delta \right)$$

para $l > 0$.

Del teorema anterior podemos deducir la prima pura de riesgo (PPR) conjunta como el valor esperado de función de distribución de la pérdida póliza conjunta:

$$PPR = E(L) = \int_0^{\infty} l \times f_L(l | \mu, \delta, \lambda, \theta) dl$$

Esta expresión es la que utilizaremos en el presente trabajo de investigación para el cálculo de la prima pura de riesgo conjunta y que recoge el efecto de dependencia entre la frecuencia y la severidad.

2.2.4.5 Modelo de regresión para la frecuencia y severidad con cópulas

La metodología descrita en las secciones anteriores muestra como modelar la frecuencia y severidad de manera conjunta sin la presencia de covariables.

En esta sección describimos como incluir covariables para estimar los parámetros de la severidad y la frecuencia, en presencia de dependencia. Para incluir covariables en nuestro análisis utilizamos el enfoque desarrollado en (Czado et al., 2012). Para ello extendemos el modelo conjunto visto en la Definición 6, permitiendo que las distribuciones marginales del costo X y el número de siniestros Y dependan de un conjunto de covariables. **Más precisamente, aplicamos los modelos lineales generalizados para los problemas de regresión marginal y los combinamos con familias de cópulas bivariadas.**

Formulación del modelo: Sean $X_i \in \mathbb{R}_+$, $i = 1, 2, 3, \dots, n$ variables aleatorias independientes continuas e idénticamente distribuidas con distribución Gamma, y sean $Y_i \in \mathbb{N}_{\geq 0}$, $i = 1, 2, 3, \dots, n$ variables aleatorias independientes discretas idénticamente distribuidas con distribución de Poisson. Modelamos X_i en términos de un vector de covariables $r_i \in \mathbb{R}^p$ e Y_i en términos de un vector de covariables $s_i \in \mathbb{R}^q$. Los modelos de regresión marginales son especificados de la siguiente manera

$$\begin{aligned} X_i &\sim \text{Gamma}(\mu_i, \delta) \quad \text{donde } \ln(\mu_i) = r_i^\top \alpha \\ Y_i &\sim \text{Poisson}(\lambda) \quad \text{donde } \ln(\lambda_i) = \ln(e_i) + s_i^\top \beta \end{aligned}$$

el componente e_i denota a la exposición medida en años póliza.

Estimación de parámetro:

El objetivo es estimar el siguiente vector parámetro desconocido:

$$v := (\alpha^\top, \beta^\top, \theta, \delta)^\top \in \mathbb{R}^{p+q+2}$$

basados en n pares de observación (x_i, y_i) . Utilizando el método de máxima verosimilitud visto en la subsección 2.2.1.5, se define la función *log-likelihood* de parámetros como:

$$l(v | x, y) = \sum_{i=1}^n \ln(f_{X,Y}(x_i, y_i | v))$$

siendo

$$x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n \quad \text{y} \quad y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$$

La estimación de máxima verosimilitud está dada por

$$\hat{v} = \operatorname{argmax} l(v | x, y)$$

En general, no existe una solución de forma cerrada para el problema de optimización. Por lo tanto, tenemos que maximizar el *log-likelihood* recurriendo a métodos numéricos. En este trabajo usaremos el algoritmo de optimización BFGS (un método cuasi Newton) para maximizar el *log-likelihood*, que viene ya implementado en el software R. Como el parámetro de la cópula $\theta \in \Theta$ es en general restricta, como visto en la Tabla 2.5, transformamos θ via una función $g : \Theta \rightarrow \mathbb{R}$ tal que $g(\theta)$ sea no restricta. En la Tabla 2.8 mostramos las transformaciones que usaremos para las cópulas en este trabajo.

TABLA N° 2.8: Trasformación no restricta de θ

Familia	Rango de θ	$g(\theta)$
Gauss	$] -1, 1[$	$\frac{1}{2} \log\left(\frac{1+\theta}{1-\theta}\right)$
Clayton	$] 0, \infty[$	$\log(\theta)$
Gumbel	$[1, \infty[$	$\log(\theta - 0.9999)$
Frank	$\mathbb{R} - \{0\}$	$\max(\theta, 0.0001 \cdot \operatorname{signo}(\theta))$

Se procede a optimizar el *log-likelihood* con respecto al vector de parámetros

$$(\alpha^\top, \beta^\top, g(\theta), \delta)^\top$$

Otra alternativa para estimar el vector de parámetros del modelo de regresión basado en cópulas, es aplicando el principio IFM (IFM por sus siglas en inglés, inference function for margins). El IFM es un procedimiento de estimación en dos pasos, propuesto por [Joe \(2014\)](#), un primer paso para estimar los parámetros de las marginales y el segundo paso para estimar el parámetro asociado a la cópula.

- Estimamos los modelos de regresión marginal de la frecuencia y de la severidad, mediante estimación máxima verosímil. Entonces calculamos por cada observación:

$$\begin{aligned}\hat{\mu} &= \exp(R\hat{\alpha}) \in \mathbb{R}^n \\ \hat{\lambda} &= \exp(S\hat{\alpha}) \odot e \in \mathbb{R}^n\end{aligned}$$

y una estimación de $\hat{\delta} \in \mathbb{R}$ como el parámetro de dispersión. En el modelo de la frecuencia, e representa la variable *offset* que es la exposición medida en años póliza, y \odot denota una multiplicación por elementos de dos vectores. Los parámetros estimados son usados para transformar las observaciones x e y en valores que serán evaluados en la cópula posteriormente,

$$\begin{aligned}u_i &:= F_X(x_i | \hat{\mu}_i, \hat{\delta}) \in [0, 1] \\ v_i &:= F_X(y_i | \hat{\lambda}_i) \in [0, 1] \\ w_i &:= F_Y(y_i - 1 | \hat{\lambda}_i) \in [0, 1]\end{aligned}$$

donde F_X y F_Y son la funciones de distribución Gamma y Poisson, respectivamente.

- Luego optimizamos el parámetro cópula θ maximizando el *log-likelihood*

$$\tilde{l}(\theta | u, v) := \sum_{i=1}^n \ln(C_1(u_i, v_i | \theta) - C_1(u_i, w_i | \theta))$$

La función \tilde{l} puede ser maximizada numéricamente. En general, el tiempo de ejecución del método IFM es mucho menor comparado con el método MLE. Además, en [Joe \(2014\)](#), el autor demuestra que los estimadores que da método IFM son consistentes y asintóticamente normal.

2.2.4.6 Distribución asintótica de los parámetros de regresión

Para la construcción de los intervalos de confianza, se realiza mediante la Matriz de Información de Fisher que se define como

$$\mathcal{J}(v) := E \left[\frac{\partial l(v | x, y)}{\partial v} \cdot \left(\frac{\partial l(v | x, y)}{\partial v} \right)^{\top} \right] \in \mathbb{R}^{(p+q+2) \times (p+q+2)}$$

Bajo condiciones de regularidad se demuestra que

$$\sqrt{n}(v - \hat{v}) \xrightarrow{D} \mathcal{N}_{p+q+2}(0, \mathcal{J}^{-1}(v))$$

donde, \mathcal{N}_k denota una distribución normal multivariante de dimensión k . Para la estimación de la información de Fisher, utilizamos el hecho de que

$$\mathcal{J}(v) = -E \left[\frac{\partial^2 l(v|x,y)}{\partial^2 v} \right]$$

y usamos la Matriz de Información de Fisher

$$\hat{\mathcal{J}}(v) := -\frac{\partial^2 l(v|x,y)}{\partial^2 v}$$

Lo anterior es la matriz Hessiana de la función *log-likelihood*. En nuestro caso, es factible calcular explícitamente las segundas derivadas parciales. Además, el algoritmo de optimización BFGS devuelve una aproximación de la matriz Hessiana que se obtiene mediante derivadas numéricas (Joe, 2014; Rao et al., 2010).

2.2.5 Construcción de modelos multivariados altamente dependientes - 2 Alternativas

El procedimiento dado por el método IFM es para un vector bivariado. Cuando se tiene dimensiones mayores el procedimiento es mucho más complejo. En esta sección presentamos dos alternativas para construir modelos multivariados con dependencia.

2.2.5.1 Alternativa 1: Construcción de Distribuciones Conjuntas Multivariantes

Considere el vector $X = (X_1, \dots, X_n)$ de variables aleatorias con función de densidad conjunta $f(x_1, \dots, x_n)$. Esta densidad puede ser factorizada como

$$f(x_1, \dots, x_n) = f(x_n) \cdot f(x_{n-1}|x_n) \cdot f(x_{n-2}|x_{n-1}, x_n) \cdots f(x_1|x_2, x_3, \dots, x_n)$$

esta descomposición es única.

Si la distribución marginal de cada variable es $F_1(x_1), \dots, F_n(x_n)$, y de acuerdo al Teorema de Sklar, la función de distribución conjunta puede ser escrito como:

$$F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n))$$

para alguna cópula C de dimensión n . A su vez, la cópula C se puede expresar de la siguiente manera

$$C(u_1, \dots, u_n) = F(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n))$$

donde $F_i^{-1}(u_i)$ es la función de distribución inversa de las marginales.

En [Aas et al. \(2009\)](#), los autores demuestran que la función de densidad conjunta f se puede escribir de la siguiente manera:

$$f(x_1, \dots, x_n) = c_{1\dots n}(F_1(x_1), \dots, F_n(x_n)) \times f_1(x_1) \times f_2(x_2) \times \dots \times f_n(x_n)$$

donde $c_{1\dots n}(\cdot)$ es la densidad de la función cópula n-variada.

Para una mejor comprensión de cómo podemos obtener la función de densidad conjunta multivariada, presentamos el caso para la función de densidad de 3 dimensiones. Esta tiene 6 maneras de poder descomponerse usando probabilidades condicionales:

$$\begin{aligned} f(x_1, x_2, x_3) &= f_{1|2,3}(x_1|x_2, x_3) f_{2|3}(x_2|x_3) f_3(x_3) \\ &= f_{1|2,3}(x_1|x_2, x_3) f_{3|2}(x_3|x_2) f_2(x_2) \\ &= f_{2|1,3}(x_2|x_1, x_3) f_{1|3}(x_1|x_3) f_1(x_1) \\ &= f_{2|1,3}(x_2|x_1, x_3) f_{3|1}(x_3|x_1) f_3(x_3) \\ &= f_{3|1,2}(x_3|x_1, x_2) f_{1|2}(x_1|x_2) f_1(x_1) \\ &= f_{3|1,2}(x_3|x_1, x_2) f_{2|1}(x_2|x_1) f_2(x_2) \end{aligned}$$

Si sólo nos enfocamos en una sola combinación, por ejemplo la primera

$$f_{1|2,3}(x_1|x_2, x_3) f_{2|3}(x_2|x_3) f_3(x_3)$$

y la escribimos en términos de cópulas bivariadas y densidades marginales:

$$f_{2|3}(x_2|x_3) = c_{2,3}(F_2(x_2), F_3(x_3)) f_2(x_2)$$

$$f_{1|2,3}(x_1|x_2, x_3) = c_{1,2|3}(F_{1|3}(x_1|x_3), F_{2|3}(x_2|x_3)) f_{1|3}(x_1|x_3)$$

el término $f_{1|3}(x_1|x_3)$ se puede descomponer aún más como

$$f_{1|3}(x_1|x_3) = c_{1,3}(F_1(x_1), F_3(x_3)) f_1(x_1)$$

tomando en cuenta los resultados parciales obtenemos el siguiente resultado

$$\begin{aligned} f(x_1, x_2, x_3) &= f_1(x_1) f_2(x_2) f_3(x_3) \\ &\quad \times c_{2,3}(F_2(x_2), F_3(x_3)) c_{1,3}(F_1(x_1), F_3(x_3)) \\ &\quad \times c_{1,2|3}(F_{1|3}(x_1|x_3), F_{2|3}(x_2|x_3)) \end{aligned}$$

Resulta que esta descomposición se puede visualizar muy bien en una estructura gráfica, como lo sugiere [Aas et al. \(2009\)](#).

Vine Cópulas: Los Vines cópulas son modelos gráficos flexibles para describir cópulas

multivariantes usando una cascada de cópulas bivariadas, llamadas *par-cópulas*.

Toda función de densidad $f(x_1, \dots, x_n)$ puede ser escrita de la siguiente manera condicional y con funciones cópulas:

$$f(x_i | v) = c_{x_i v_i | v_{-j}} (F(x_i |, F(v_i | v_{-j}))) F(x_i | v_{-j})$$

donde $v = \{x_{i+1}, \dots, x_n\}$ es el conjunto condicional de las distribuciones marginales de x_i , la variable $v_i \in v$ y v_{-j} es el conjunto de variables en v después de extraer v_i con $i = 1, \dots, n-1$ y $c(u_1, u_2)$ es la cópula de densidad definida como $\frac{\partial c(u_1, u_2)}{\partial u_1 \partial u_2}$.

Cuando $f(x_i | v)$ es iterativamente descompuesta, ésta se convierte en el producto bivariado de densidad de cópulas y funciones marginales de x_i . si estas marginales también son descompuestas de forma iterativa, entonces $f(x_1, \dots, x_n)$ es el producto de densidad de cópulas bivariadas y sus marginales. Dos especiales *pair – copulas* son utilizadas y cuyas densidades son las siguientes:

C-Vine Cúpula

$$f(x_1, \dots, x_n) = \prod_{k=1}^n f_k(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j, i+1 | 1, \dots, j-1} (F_{j|1, \dots, j-1}(x_j | x_{1, \dots, j-1}), F_{j+i|1, \dots, j-1}(x_{j+i} | x_{1, \dots, j-1}))$$

D-Vine Cúpula

$$f(x_1, \dots, x_n) = \prod_{k=1}^n f_k(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{j, i+j | i+1, \dots, i+j-1} (F_{j|i+1, \dots, i+j-1}(x_j | x_{i+1, \dots, i+j-1}), F_{j+i|i+1, \dots, i+j-1}(x_{j+i} | x_{i+1, \dots, i+j-1}))$$

Como se indicó, estas estructuras Vine pueden ser más fácilmente entendidas por medio de gráficos utilizando árboles.

Por ejemplo, para el caso de una función de densidad conjunta de 5 variables $f(x_1, x_2, x_3, x_4, x_5)$ seguimos los *pair – copulas* generados en el arbol del D-vine cúpula para la construcción de la densidad conjunta:

Densidad conjunta utilizando D-vine cúpula:

$$\begin{aligned} f(x_1, x_2, x_3, x_4, x_5) &= f_1(x_1) \times f_2(x_2) \times f_3(x_3) \times f_4(x_4) \times f_5(x_5) \\ &\times c_{12}(F_1(x_1), F_2(x_2)) \times c_{23}(F_2(x_2), F_3(x_3)) \times c_{34}(F_3(x_3), F_4(x_4)) \times c_{45}(F_4(x_4), F_5(x_5)) \\ &\times c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \times c_{24|3}(F_{2|4}(x_2|x_4), F_{2|3}(x_2|x_3)) \times c_{35|4}(F_{3|4}(x_3|x_4), F_{5|4}(x_5|x_4)) \\ &\times c_{14|23}(F_{1|23}(x_1|x_2, x_3), F_{4|23}(x_4|x_2, x_3)) \times c_{25|34}(F_{2|34}(x_2|x_3, x_4), F_{5|34}(x_5|x_3, x_4)) \\ &\times c_{15|234}(F_{1|234}(x_1|x_2, x_3, x_4), F_{5|234}(x_5|x_2, x_3, x_4)) \end{aligned}$$

FIGURA N° 2.14: D-vine cópula: 5 variables, 4 árboles y 10 pair-cópulas

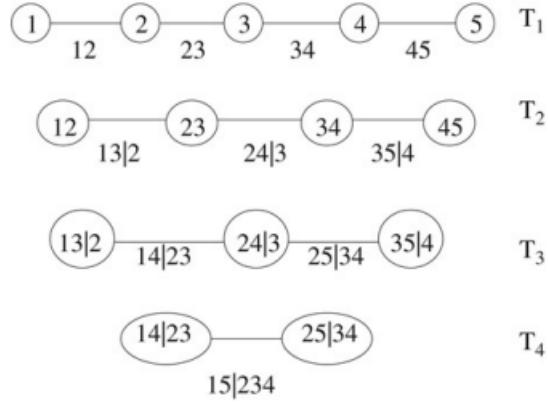
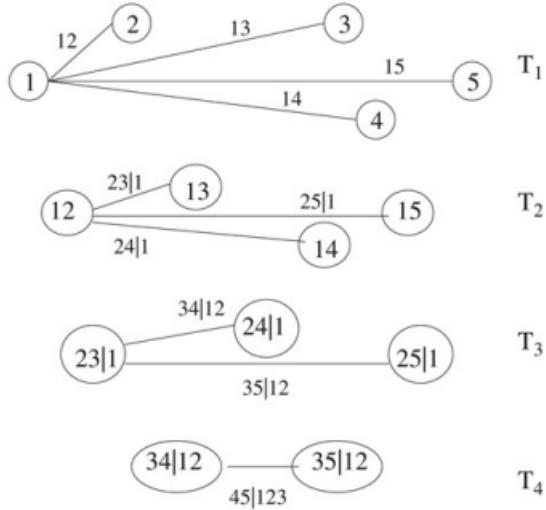


FIGURA N° 2.15: C-vine cópula: 5 variables, 4 árboles y 10 pair-cópulas



Densidad conjunta utilizando C-vine cópula:

$$\begin{aligned}
 f(x_1, x_2, x_3, x_4, x_5) &= f_1(x_1) \times f_2(x_2) \times f_3(x_3) \times f_4(x_4) \times f_5(x_5) \\
 &\times c_{12}(F_1(x_1), F_2(x_2)) \times c_{13}(F_1(x_1), F_3(x_3)) \times c_{15}(F_1(x_1), F_5(x_5)) \times c_{14}(F_1(x_1), F_4(x_4)) \\
 &\times c_{23|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1)) \times c_{25|1}(F_{2|1}(x_2|x_1), F_{5|1}(x_5|x_1)) \times c_{24|1}(F_{2|1}(x_2|x_1), F_{4|1}(x_4|x_1)) \\
 &\times c_{34|12}(F_{3|12}(x_3|x_1, x_2), F_{4|12}(x_4|x_1, x_2)) \times c_{35|12}(F_{3|12}(x_3|x_1, x_2), F_{5|12}(x_5|x_1, x_2)) \\
 &\times c_{45|123}(F_{4|123}(x_4|x_1, x_2, x_3), F_{5|123}(x_5|x_1, x_2, x_3))
 \end{aligned}$$

2.2.5.2 Alternativa 2: Agregación jerárquica de riesgos basado en cópula

Para fines de agregación de riesgos de alta dimensión, las clases de cópula más populares son demasiado restrictivas en términos de estructuras de dependencia alcanzables. Estas limitaciones se agravan a medida que aumentan las dimensiones. En **Arbenz, Hummel, y**

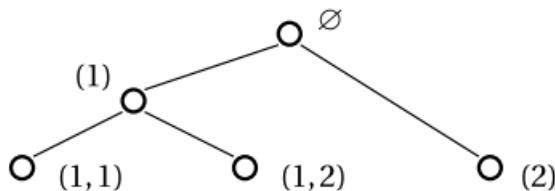
Mainik (2012) los autores plantean un método de agregación de riesgos jerárquico que es flexible en dimensiones altas. Nosotros seguiremos este trabajo para describir esta sección. Con este método basta con especificar una cópula de baja dimensión para cada paso de agregación en la jerarquía. Se pueden combinar cópulas y márgenes de cualquier tipo.

Definición 8. Un conjunto finito $\tau \subset \emptyset \cup \bigcup_{n=1}^{\infty} \mathbb{N}^n$ denota un árbol enraizado si:

1. La raíz \emptyset está contenida en τ .
2. para cada nodo $I = (i_1, \dots, i_n) \in \tau$ el número de nodos hijos está dado por $N_I \in \mathbb{N}_0$, es decir, el nodo I tiene un hijo $(i_1, \dots, i_n, k) \in \tau$ si y solo si $k \in \{n \in \mathbb{N} : 1 \leq n \leq N_I\}$,
3. cada nodo $(i_1, \dots, i_d) \in \tau$ tiene un nodo pariente representado por $(i_1, \dots, i_d) \in \tau$.

Ejemplo: Sea $\tau = \{\emptyset, (1), (1, 1), (1, 2), (2)\}$, como es ilustrado en la Figura N° 2.16, los nodos $(1, 1)$, $(1, 2)$ y (2) son nodos hojas. Los nodos \emptyset y (1) son nodos ramas. Tenemos $N_{\emptyset} = N_1 = 2$ y $N_{1,1} = N_{1,2} = N_2 = 0$

FIGURA N° 2.16: Ilustración del árbol $\tau = \{\emptyset, (1), (1, 1), (1, 2), (2)\}$



Árboles de agregación por dependencia: Esta sección presenta un enfoque de agregación de riesgos basado en un árbol enraizado dado τ . Definimos sobre algún espacio de probabilidad $(\Omega, \mathcal{A}, \mathbb{P})$, un vector aleatorio $(X_I)_{I \in \tau}$ que asigna a cada nodo $I \in \tau$ una variable aleatoria $X_I : \Omega \rightarrow \mathbb{R}$ tal que:

- Para cada nodo hoja $I \in \mathcal{L}(\tau)$, la variable X_I representan los riesgos cuyo agregado nos interesa,
- Los nodos ramas, $X_I, I \in \mathcal{B}(\tau)$, están dados por la agregación de los su hijos: $X_I = X_{I,1} + \dots + X_{I,N_I}$

El enfoque de modelado consiste en definir distribuciones marginales $F_I, I \in \mathcal{L}(\tau)$ y las estructuras de dependencia para los pasos de agregación de modo que podamos calcular la distribución del agregado total X_{\emptyset} . Tenga en cuenta que por inducción

$$X_{\emptyset} = X_1 + \dots + X_{N_{\emptyset}} = \sum_{I \in \mathcal{L}(\tau)} X_I$$

Con este enfoque podemos evitar modelar explícitamente la cópula de alta dimensión que describe la dependencia entre todos los nodos hoja $X_I, I \in \mathcal{L}(\tau)$. Para facilitar la notación,

eliminamos los corchetes de los vectores índice, así como el argumento τ siempre que el significado sea claro. Por ejemplo, $X_{(1,1)} = X_{1,1}$, $\mathcal{D}(I, \tau) = \mathcal{D}(I)$ y $\mathcal{LD}((1,1), \tau) = \mathcal{LD}(1,1)$

Sea la siguiente tripleta:

$$\left(\tau, (F_I)_{I \in \mathcal{L}(\tau)}, (C_I)_{I \in \mathcal{B}(\tau)} \right) \quad (2.16)$$

que consiste en:

- un árbol enraizado τ ;
- Funciones de distribución $F_I : \mathbb{R} \rightarrow [0, 1]$ para todo $I \in \mathcal{L}(\tau)$
- cópula $C_I : [0, 1]^{\mathbb{N}_I} \rightarrow [0, 1]$ para todo $I \in \mathcal{B}(\tau)$

Por ejemplo, en el ejemplo anterior podemos identificar a $\tau = \{\emptyset, (1), (1,1), (1,2), (2)\}$, tres distribuciones univariadas $F_{1,1}, F_{1,2}$ y F_2 y dos cópulas bivariadas C_1 y C_\emptyset .

Definición 9. Dada una tripleta como en (2.16), $(X_I)_{I \in \tau}$ es llamada ligeramente dependiente de los árboles si las siguientes condiciones son conocidas:

- Para cada nodo hoja $I \in \mathcal{L}$, la variable aleatoria X_I tiene una distribución $F_I : \mathbb{P}[X_I \leq x] = F_I(x)$ para todo $x \in \mathbb{R}$.
- Para cada nodo rama $I \in \mathcal{B}$, X_I es la suma de sus hijos, es decir, $X_I = \sum_{i=1}^{N_I} X_{I,i}$. La distribución de probabilidad marginal acumulada de X_I está denotada por $F_I : \mathbb{R} \rightarrow [0, 1]$.
- Para cada nodo rama $I \in \mathcal{B}$, la estructura de dependencia de sus hijos $\mathcal{C}(I)$ está dada por la cópula C_I , es decir,

$$\mathbb{P}[X_{I,1} \leq x_1, \dots, X_{I,N_I} \leq x_{N_I}] = C_I(F_{I,1}(x_1), \dots, F_{I,N_I}(x_{N_I}))$$

, para todo $(x_1, \dots, x_{N_I}) \in \mathbb{R}^{N_I}$

A continuación mostramos como se usa la agregación jerárquica de riesgos basado en cópula con un ejemplo explícito.

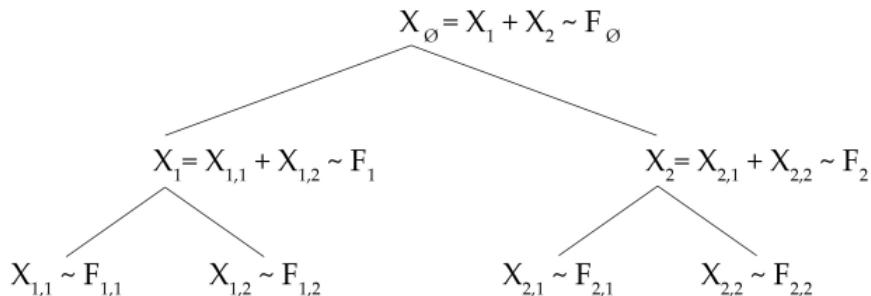
Ejemplo: Supongamos que se nos dan cuatro riesgos diferentes en un vector aleatorio $X = (X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2})$. Aquí $X_{1,1}$ significa “Seguro de automóvil en Suiza” y $X_{1,2}$ significa “Seguro de automóvil en Italia”. Además, $X_{2,1}$ significa “Terremoto Suiza” y $X_{2,2}$ significa “Terremoto Italia”. En caso de que estemos interesados en el riesgo agregado $X_\emptyset := X_{1,1} + X_{1,2} + X_{2,1} + X_{2,2}$, podríamos intentar encontrar un modelo para la función de distribución conjunta F de X modelando primero las distribuciones marginales. $F_{1,1}, F_{1,2}, F_{2,1}$ y $F_{2,2}$ del riesgo individual e imponer un cópula C entre el riesgo individual. La distribución conjunta estaría entonces dada por

$$F(x_1, x_2, x_3, x_4) = C(F_{1,1}(x_1), F_{1,2}(x_2), F_{2,1}(x_3), F_{2,2}(x_4))$$

y la distribución de X_\emptyset se puede calcular directamente a partir de F . Alternativamente, en un primer paso podríamos modelar las distribuciones de los riesgos parciales, $(X_{1,1}, X_{1,2})$ y $(X_{2,1}, X_{2,2})$, mediante combinando los riesgos $X_{1,1}$ y $X_{1,2}$ mediante una cópula bivariada C_1 , mientras que $X_{2,1}$ y $X_{2,2}$ se combinan mediante una cópula bivariada C_2 . Conociendo las distribuciones de los riesgos parciales, podemos calcular fácilmente la distribución de las sumas parciales $X_1 := X_{1,1} + X_{1,2}$ y $X_2 := X_{2,1} + X_{2,2}$. Las sumas parciales, X_1 y X_2 , pueden luego combinarse nuevamente mediante un modelo de cópula bivariado adecuado C_\emptyset . Finalmente, esto nos permite calcular la distribución F_\emptyset del agregado total $X_\emptyset := X_1 + X_2 = X_{1,1} + X_{1,2} + X_{2,1} + X_{2,2}$.

Este procedimiento se conoce como “agregación jerárquica de riesgos basada en cópulas” y se ilustra mejor mediante el llamado “árbol de agregación”. El árbol de agregación representa gráficamente de qué manera están vinculados los riesgos individuales y qué estructura de dependencia se supone que existe entre ellos. La [Figura N° 2.17](#) muestra el árbol de agregaciones correspondiente a la situación descrita anteriormente.

FIGURA N° 2.17: Una ilustración de un modelo de agregación de 4 dimensiones



La principal ventaja de la agregación jerárquica de riesgos es que no necesitamos especificar la cópula de todos los riesgos. Una vez más enfatizamos que es extremadamente improbable encontrar un modelo de cópula que describa adecuadamente la estructura de dependencia entre un gran número de riesgos. Las observaciones conjuntas entre todos los riesgos son demasiado raras y las estructuras de dependencia alcanzables de los modelos de cópula paramétricos comunes son demasiado limitadas. En lugar de ello, agregamos los riesgos jerárquicamente, lo que requiere únicamente especificar la dependencia conjunta entre las subcarteras agregadas en los diferentes pasos de agregación. Para estas subcarteras podemos utilizar cópulas de baja dimensión, que son bien conocidas por ser medidas de dependencia más realistas y ofrecer mayor flexibilidad. En particular, este modelo permite utilizar diferentes características de dependencia (dependencia de cola, asimetría radial, etc.) para cada paso de agregación.

El modelo puede verse como una herramienta de reducción de dimensiones: la complejidad y el número de parámetros se pueden ajustar a la complejidad necesaria y a la información disponible. Como consecuencia lógica, el método no especifica la distribución conjunta

completa de los riesgos individuales. En otras palabras, en general hay más de una distribución que satisface un modelo de árbol de agregación determinado. Llamaremos a esas distribuciones “ligeramente dependientes de los árboles”.

2.3 Enfoque teórico asumido por el investigador

2.3.1 Prima pura sujeto a efectos de dependencia

En este trabajo se presentará una propuesta metodológica alternativa para el cálculo de la prima pura de riesgo, si bien se utilizará modelos lineales generalizados para modelar la frecuencia y la severidad en presencia de covariables, una herramienta de amplio uso en la industria aseguradora, mi propuesta se diferencia en el sentido que considero el grado de dependencia entre la frecuencia y la severidad. Además, las primas calculadas por cobertura serán agregadas considerando el grado de asociación que estas tienen, ya que en mi experiencia una cobertura activada suele estar acompañada de otras o tal vez ninguna al momento de la ocurrencia del siniestro. Agregarlas bajo una suma simple nos hace intuir rápidamente que existe una sobreestimación en las primas de riesgo calculadas.

2.3.1.1 Formulación metodológica

Para explicar de forma sencilla el modelo de prima de pura de riesgo (PPR) sujeto a dependencia, el modelo consta de los siguientes pasos:

1. Primero calculamos los modelos marginales de regresión generalizado de la frecuencia y de la severidad, esto por cada cobertura.
2. Las coberturas que se considerarán en este trabajo son las aplicadas en un seguro de automóviles:
 - PP: Pérdidas Parciales
 - PT: Pérdidas Totales
 - RC: Responsabilidad Civil
 - AS: Asistencias

Considérese que estas coberturas hacen más de 95 % de la prima. En una venta real de estos seguros suelen ofrecerse muchas coberturas en productos de ”Todo Riesgo”.

3. Luego calculamos la densidad conjunta, en el caso de pérdidas parciales la densidad es $f_{PP}(X, N|\theta)$.
4. Con la funciones de densidad conjunta calculamos el valor esperado, que será la prima

pura de riesgo. En el caso de las pérdidas parciales tendremos

$$PPR_{PP} = \int_{l=0}^{\infty} l \times f_{PP}(X, N|\theta)$$

donde l representa a la pérdidas incurridas por cobertura de pérdidas parciales.

5. Una vez calculadas las primas puras de riesgo por coberturas procedemos a la agregación bivariada de las mismas mediante funciones cópulas. La prima pura de riesgo agregada de las coberturas Pérdida Parcial y Pérdida Total es

$$PPR_{PP+PT|\theta} = PPR_{PP} \oplus PPR_{PT}$$

donde el operador \oplus es definido como una suma de riesgos con dependencia.

6. Por último, agregamos las primas bivariadas para obtener la prima total por las cuatro coberturas

$$PPR_{PP+PT+RC+AS|\theta} = (PPR_{PP} \oplus PPR_{PT}) \oplus (PPR_{RC} \oplus PPR_{AS})$$

En resumen el modelo propuesto tiene la siguiente forma esquemática, que va de derecha (modelos marginales) a izquierda (Prima Pura de riesgo conjunta total)

$$PPR_{PP+PT+RC+AS|\theta_7} \left\{ \begin{array}{l} PPR_{PP+PT|\theta_5} \left\{ \begin{array}{l} f_{PP}(X, N|\theta_1) \left\{ \begin{array}{l} f_{PP}(N_{PP}) : \text{Modelo Marginal de la frecuencia} \\ f_{PP}(X_{PP}) : \text{Modelo Marginal de la Severidad} \end{array} \right. \\ f_{PT}(X, N|\theta_2) \left\{ \begin{array}{l} f_{PT}(N_{PT}) : \text{Modelo Marginal de la frecuencia} \\ f_{PT}(X_{PT}) : \text{Modelo Marginal de la Severidad} \end{array} \right. \end{array} \right. \\ PPR_{RC+AS|\theta_6} \left\{ \begin{array}{l} f_{RC}(X, N|\theta_3) \left\{ \begin{array}{l} f_{RC}(N_{RC}) : \text{Modelo Marginal de la frecuencia} \\ f_{RC}(X_{RC}) : \text{Modelo Marginal de la Severidad} \end{array} \right. \\ f_{AS}(X, N|\theta_4) \left\{ \begin{array}{l} f_{AS}(N_{AS}) : \text{Modelo Marginal de la frecuencia} \\ f_{AS}(X_{AS}) : \text{Modelo Marginal de la Severidad} \end{array} \right. \end{array} \right. \end{array} \right. \right.$$

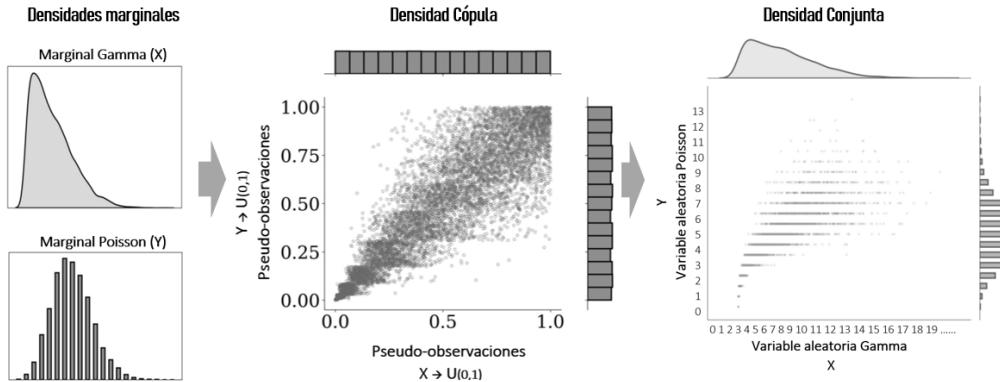
donde θ_i es el parámetro de dependencia de la cópula correspondiente.

Para las densidades conjuntas de la frecuencia y de la severidad el proceso es como el descrito en el punto 2.2.4.3, sin embargo en la [Figura N° 2.18](#) podemos observar el proceso de modelación

Para la elección de la mejor función cópula, compararemos los *log-likelihoods* de cada modelo, el máximo de ellos no dirá qué familia de cópula elegir.

Una vez definido la función de densidad conjunta, calculamos la prima pura de riesgo como

FIGURA N° 2.18: Densidad conjunta de la Frecuencia y Severidad



su valor esperado. Para ello utilizamos la fórmula definida en el Teorema 10.

Para realizar la suma de riesgos X e Y sujetos a dependencia, debemos conocer su distribución de probabilidad, en nuestro caso ambos son Gamma, dado que tanto X como Y representan al conjunto de primas de riesgos por coberturas. Con la data disponible (x_i, y_i) estimamos sus parámetros mediante máxima verosimilitud y procedemos a realizar los siguientes pasos:

- Generamos la primera variable aleatoria \tilde{X} con distribución F_X , es decir:

$$u \sim U(0,1) \rightarrow \tilde{X} = F_X(u)$$

- Generamos la segunda variable aleatoria \tilde{Y} con distribución F_Y , es decir:

$$v \sim U(0,1) \rightarrow \tilde{Y} = F_Y(v)$$

- Hallamos la primera derivada de la función cúpula a partir de la densidad conjunta

$$P[X \leq \tilde{X}, Y \leq \tilde{Y}] = C(u, v)$$

y la función condicional definida como

$$P[X \leq \tilde{X}|Y \leq \tilde{Y}] = \frac{\partial}{\partial u} C(u, v) = q(u, v)$$

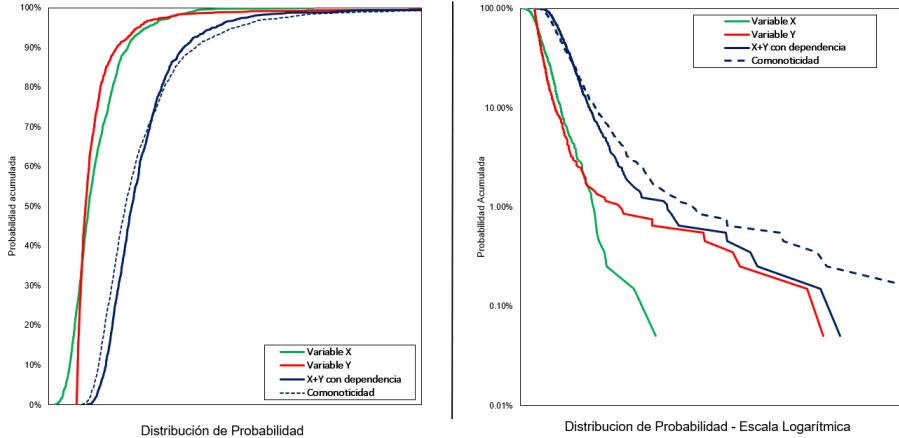
- De $q(u, v)$ despejamos v para obtener $v = v(u, q)$, con ello podemos simular $u \sim U(0,1)$ y $q \sim U(0,1)$ para obtener los valores de v .
- Para las familias de cúpulas que utilizaremos en este trabajo mostramos las expresiones de v para cada una:

TABLA N° 2.9: Expresiones de v por cada familia de cópulas

Familia	$C(u, v \theta)$	$q = \frac{\partial}{\partial u} C(u, v)$	v
Gauss	$\Phi_2(\Phi^{-1}(u), \Phi^{-1}(v) \theta)$	$\Phi\left(\frac{\Phi^{-1}(v)-\theta\Phi^{-1}(u)}{\sqrt{1-\theta^2}}\right)$	$\Phi\left(\theta\Phi^{-1}(u)+\Phi^{-1}(q)(\sqrt{1-\theta^2})\right)$
Clayton	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta-1}u^{-\theta-1}$	$(1+u^{-\theta}(q^{\frac{-\theta}{\theta+1}}-1))^{\frac{-1}{\theta}}$
Gumbel	$\exp\left(-((-\log u)^\theta + (-\log v)^\theta)^{1/\theta}\right)$	$u^{-1}\exp\left(-((-\log u)^\theta + (-\log v)^\theta)^{1/\theta}\right)$	$\exp\left(-((-\ln qu)^\theta + (-\ln u)^\theta)^{\frac{1}{\theta}}\right)$
Frank	$-\frac{1}{\theta}\log\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right)$	$\frac{e^\theta(e^{\theta v}-1)}{e^{\theta(u+1)}+e^{\theta(v+1)}-e^\theta-e^{\theta(u+v)}}$	$\frac{1}{\theta}\ln\left(\frac{qe^\theta(e^{\theta u}-1)}{e^\thetaqe^\theta(1-e^u)}\right)$

Una vez calculado u y v podemos obtener F_X y F_Y y F_{X+Y} y su correspondiente función de comonotonicidad asociada a la suma de variables aleatorias. En el presente trabajo la prima total es la agregación bivariada de las primas por cobertura que se van calculando en cada paso. Las primas por cobertura se ajustan a una función de distribución de probabilidad Gamma, esto ya se dijo antes, cuyos parámetros son estimados mediante máxima verosimilitud.

FIGURA N° 2.19: Distribución de probabilidad de riesgos agregados



Fuente: Elaboración Propia usando el software R.

A modo de ejemplo, supongamos que tenemos dos primas $X = 650$ y $Y = 21$, siendo $X \sim \text{Gamma}(\alpha = 0.98, \beta = 798)$ y $Y \sim \text{Gamma}(\alpha = 0.67, \beta = 8.6)$, supongamos una cópula Clayton de parámetro $\theta = 0.71$ para la agregación con dependencia de las primas de riesgo.

- Calculamos $u = F_X^{-1}(650) = 0.57$ y $\hat{v} = F_Y^{-1}(21) = 0.96$
- Con u y \hat{v} podemos calcular $q = C_{clayton}(u = 0.57, \hat{v} = 0.96 | \theta = 0.71) = 0.55$
- Por lo tanto, de acuerdo a la Tabla 2.9 obtenemos $v = 0.61$.
- Por lo tanto, $650 \oplus 21 = F_X^{-1}(u) + F_Y^{-1}(v) = 655$

2.4 Hipótesis

2.4.1 Hipótesis general

La propuesta metodológica del cálculo de la prima pura de riesgo sujeto a dependencia entre los riesgos de frecuencia y severidad, y agregados mediante funciones cópulas por cobertura nos permita estimaciones robustas y de acuerdo con el nivel de riesgo específico de los asegurados frente supuesto usual de independencia.

2.4.2 Hipótesis específicas

- Se puede construir un modelo de regresión conjunto de la frecuencia y de la severidad para estimar la prima pura de riesgo de una determinada cobertura.
- Se puede determinar el grado de dependencia entre la frecuencia y la severidad.
- La primas total se puede obtener mediante agregación de los riesgos asociados a cada cobertura mediante funciones cópulas.
- Se puede medir el grado de dependencia de las distribuciones de la primas de riesgo por cobertura.

2.5 Variables

2.5.1 Robustez

2.5.1.1 Definición conceptual

Se dice que una metodología es robusta respecto de las desviaciones de los supuestos del modelo, cuando el proceso continúa trabajando bien, aún cuando, en mayor o menor extensión, los supuestos no se mantienen.

2.5.1.2 Definición operacional

Para la medición de la robustez estadística del modelo GLM utilizaremos herramientas como *Simple Quantile Plots, Double Lift Charts, Loss Ratio Charts* y el Indice de Gini.

2.5.2 Asociación

2.5.2.1 Definición conceptual

La asociación estadística mide el grado de dependencia entre dos riesgos o variables, e implica una mayores aplicaciones que el coeficiente de correlación.

2.5.2.2 Definición operacional

El grado de asociación entre riesgos la mediremos mediante la τ de Kendall.

2.5.3 Desempeño

2.5.3.1 Definición conceptual

Se refiere a la comparación de los modelos candidatos, que se irán descartando mediante el uso de estadísticos que comparan el poder predictivo de los modelos frente a los datos observados.

2.5.3.2 Definición operacional

El desempeño de los modelos candidatos lo mediremos mediante el uso del *log-likelihood* y la Devianza.

CAPÍTULO III: METODOLOGÍA

3.1 Tipo, nivel y diseño de la investigación

- Tipo de investigación es teórica.
- Nivel de investigación es explicativo.
- El diseño de la investigación va ser experimental.

3.2 Población, muestra y tamaño de muestra

3.2.1 Población

Sistema estadístico del seguros privado de seguros de automóviles de Brasil, cuya información pública está disponible está desde el año 2006. Los datos se obtuvieron de <https://www2.susep.gov.br/menuestatistica/autoseg/principal.aspx>

3.2.2 Muestra

Se seleccionó la información correspondiente de emisiones de pólizas de años 2015, 2016, 2017, 2018 y 2019. Para la información de siniestros se tomaron los años 2015, 2016, 2017, 2018 y 2019. Sin embargo, para una buena estimación de la exposición, prima devengada¹ y siniestralidad se tomaron efectivamente los años 2016, 2017, 2018 y 2019.

3.2.3 Tamaño de muestra

La muestra seleccionada contiene 1,581,869 vehículos identificados por Número de póliza y número de certificado. Para los siniestros, se seleccionaron 4 coberturas: Pérdida Parcial ($N: 251,786$ y $M: 226,604,385 \text{ USD}$), Pérdida Total ($N: 6,684$ y $M: 59,175,937 \text{ USD}$), Responsabilidad Civil ($N: 24,863$ y $M: 22,872,707 \text{ USD}$) y Asistencias ($N: 14,630$ y $M: 638,490 \text{ USD}$). Donde: N es la cantidad de siniestros y M es el monto de siniestros incurridos en dólares.

3.3 Técnicas de análisis e instrumentos

3.3.1 Técnicas de Análisis

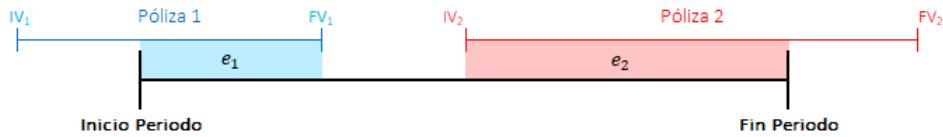
Para la estimación de la prima pura de riesgo debemos calcular de la mejor manera los siguientes indicadores actuariales que son vitales:

- **Expuestos:** Esta medida nos indica la proporción de tiempo en que un riesgo (Unidad

¹No se pudo calcular de forma adecuada debido a la falta de información para separar los impuestos de la prima comercial

vehicular) estuvo expuesto respecto de un periodo determinado, como lo indica la Figura N° 3.1.

FIGURA N° 3.1: Medida de exposición de una póliza en un periodo de tiempo



Los periodos de tiempo que utilizaremos en nuestro estudio son los años indicados en la muestra. Una fórmula de cálculo de esta medida es la siguiente:

$$0 \leq e_i = \frac{\text{Min}(\text{Fin}_{\text{periodo}}, \text{FV}_i) - \text{Max}(\text{Inicio}_{\text{periodo}}, \text{IV}_i)}{\text{Fin}_{\text{periodo}} - \text{Inicio}_{\text{periodo}}} \leq 1$$

donde i es cada unidad vehicular con Inicio de Vigencia (IV) y Fin de Vigencia (FV) de la póliza.

- **Prima Devengada:** Es la prima que se gana en el periodo expuesto. Su cálculo por unidad vehicular es de la siguiente manera:

$$\text{Prima}_{DEV} = \frac{\text{Min}(\text{Fin}_{\text{periodo}}, \text{FV}_i) - \text{Max}(\text{Inicio}_{\text{periodo}}, \text{IV}_i)}{\text{FV}_i - \text{IV}_i} \times (\text{Prima Neta})$$

donde la Prima Neta es la prima cobrada por todo el periodo de su vigencia, sin IVA² y DE³.

- **Siniestros Incurridos:** Esto monto indica el valor del siniestro, a nivel de base de datos este monto es obtenido como:

$$\text{Incurrido} = \text{Reserva Pendiente} + \text{Pagos}$$

por cada siniestro y cobertura.

- **Siniestros Incurridos Últimos:** Como en la base de datos tenemos siniestros en estado Pagados y Pendientes, esto significa que existirá aún desarrollo de los siniestros. Para poder estimar el comportamiento último de los siniestros, calculamos los factores IBNR que nos permitirá estimar los siniestros últimos.

En virtud del modelo colectivo de riesgo, por unidad vehicular debemos calcular el monto total del siniestro S_i y la cantidad total del siniestro N , es decir, por póliza debemos calcular:

$$S_i = X_1 + \dots + X_N,$$

²IVA: Impuesto a las Ventas

³DE: Derecho de Emisión

donde X_j ($1 \leq j \leq N$) es cada siniestro de la unidad vehicular i .

Las medidas básicas definidas nos permitirá realizar el cálculo de la Frecuencia y de la Severidad en términos globales, por cada cobertura. Estos valores de la Frecuencia y de la Severidad nos ayudará en la validación de los interceptos en los modelos de regresión marginales.

De acuerdo a la formulación metodológica, por cobertura debemos construir 4 modelos de regresión conjunta, por cada familia de cópula considerada en este trabajo, y de acuerdo al criterio máximo *log – likelihood* elegimos el mejor modelo.

El tratamiento de los datos se realizó utilizando gestores de base de datos como MS Access y SQL debido al alto volumen de los datos, y sobretodo por el alto performance necesarios en realizar las consultas y cruces respectivos. Las estimaciones máxima verosimiles de los parámetros de regresión y de las cópulas se realizaron en código R-Studio, utilizando los paquetes *Optim()* y *CopulaRegresion()*.⁴

3.3.2 Instrumentos

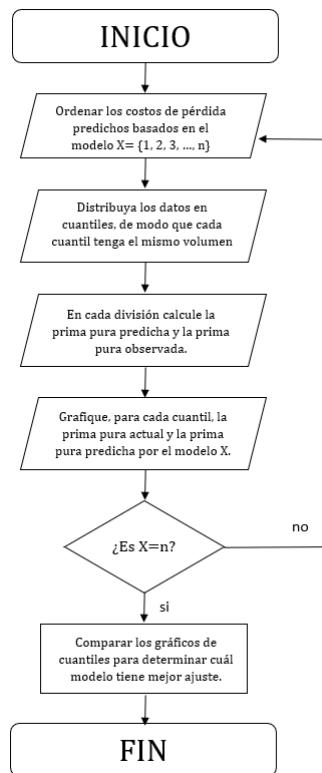
Para poder comparar los distintos modelos de regresión se podrán utilizar los siguientes indicadores:

3.3.2.1 Indicadores de Robustez

- **Simple Quantile Plots:** Estos son una representación visual sencilla de la capacidad de un modelo para diferenciar con precisión entre los mejores y los peores riesgos. Suponga que hay dos modelos, el Modelo A y el Modelo B, los cuales producen una estimación del costo de la pérdida esperada para cada asegurado. Los gráficos de cuantiles simples se crean a través de los siguientes pasos mostrados en el algoritmo de la figura 3.2. Para determinar el mejor modelo, debemos considerar 3 criterios:
 - Precisión predictiva: Se refiere a qué tan bien es capaz cada modelo predecir la prima observada en cada cuantil.
 - Monotonicidad: A medida que la prima pura predicha se incremente monótonicamente con los cuantiles, entonces la prima pura observada también se incrementará (y viceversa).
 - Distancia vertical entre el primer y último cuantil: El primer cuantil contiene los riesgos que el modelo cree tendrán una mejor experiencia, y el último cuantil contiene los riesgos que el modelo tendrá la peor experiencia. Una diferencia grande (llamada *lift*) entre la prima pura observada en los cuantiles con los costos de pérdida predichos más pequeños y más grandes indica que el modelo es capaz de distinguir al máximo los mejores y los peores riesgos.

⁴Paquete desarrollado en R por (Kholifah et al., 2019)

FIGURA N° 3.2: Algoritmo para la construcción del Simple Quantile Plot



- **Double lift Charts:** Este gráfico es similar *Simple Quantile Plots*, pero compara directamente dos modelos. Suponga que hay dos modelos, el Modelo A y el Modelo B, los cuales producen una estimación del costo de la pérdida esperada para cada asegurado. Este gráfico se crea a través de los siguientes pasos:

- Calculamos por registro

$$\text{Sort ratio} = \frac{\text{Predicted cost from model A}}{\text{Predicted cost from model B}}$$

- Ordenar los datos en función al Sort Ratio.
- Dividir los datos en cuantiles, tales como quíntiles o deciles.
- Dentro de cada división, calcular la prima pura predicha promedio del modelo A, del modelo B y de los observados. Para cada una de esas cantidades, divida el promedio del cuantil por el promedio general.
- Por cada cuantil, grafique los tres primas puras mencionadas anteriormente.

En este gráfico, el modelo “ganador” es el que se acerca más a la prima pura real en cada cuantil.

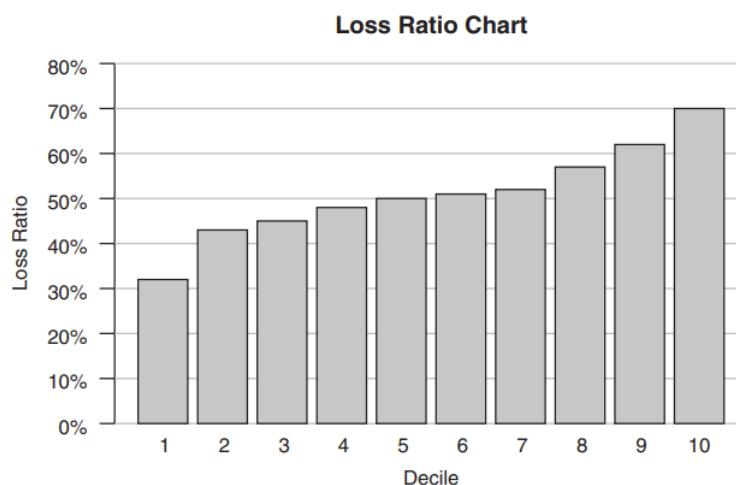
- **Loss Ratio Chart:** Este gráfico, en lugar de trazar la prima pura para cada categoría, se representa la tasa de siniestralidad. Los pasos para crear un gráfico de tasa de

pérdida son muy similares a los de crear un *Simple Quantile Plots*, con una diferencia importante:

- Ordenar los datos basados en las predicciones del modelo.
- Dividir los datos en cuantiles, de tal manera que cada cuantil tenga el mismo volumen de exposición.
- Dentro de cada división, calcular el *loss ratio* de los datos observados para los riesgos dentro la división.

Idealmente, el modelo es capaz de identificar deficiencias en el *scoring* actual al segmentar los riesgos en función del *loss ratio*. Por ejemplo, de acuerdo a la Figura 3.3.,

FIGURA N° 3.3: Ejemplo del *loss ratio chart*



podemos indicar que si un plan de scoring es perfecto, entonces todos los riesgos deberían tener el mismo *loss ratio*. El hecho de que este modelo sea capaz de segmentar los datos en divisiones de altos y bajos *loss ratios* es un fuerte indicador que el plan de scoring actual. La ventaja de este indicador es que es fácil de entender y explicar. Tener en cuenta que el *loss ratio* es comúnmente utilizado para medir la rentabilidad de la compañía de seguros.

- **El Índice de GINI:** Llamado así por el estadístico y sociólogo Corrado Gini, y es comúnmente usado en economía para cuantificar la desigualdad de ingresos en una nación. El índice de Gini puede también ser utilizado para medir el *lift* de un plan de calificación de seguros mediante la cuantificación de su capacidad para segmentar a la población en los mejores y peores riesgos. El índice de Gini para un modelo de predicción se calcula de la siguiente manera:

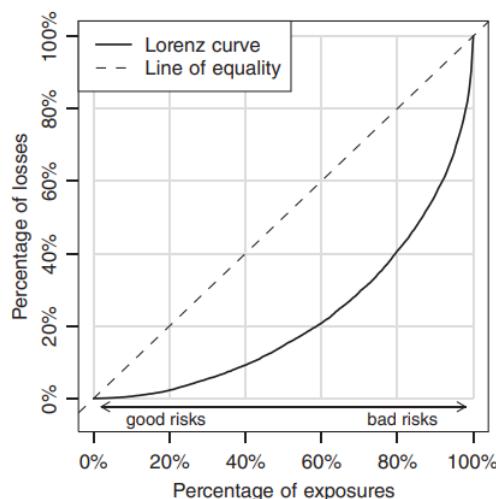
- Ordenar los datos basados en el costo de pérdida predicho por el modelo. Los registros en la parte superior de los datos son entonces los riesgos en los cuales el

modelo cree son los mejores riesgos, y los registros al último de los datos son los riesgos el cual el modelo cree son los peores riesgos.

- En el eje x , grafique el porcentaje acumulado de los expuestos.
- En el eje y , grafique el porcentaje acumulado de las pérdidas.

El lugar geométrico de los puntos es la curva de Lorenz, y el índice de Gini es el doble del área entre la curva de Lorenz y la línea de igualdad.

FIGURA N° 3.4: Índice de Gini para un Modelo de Prima Pura



De la Figura N° 3.4, el ejemplo, identifica el 60% de los expuestos el cual contribuye solo el 20% del total de pérdidas. Su índice de GINI es del 56.1%. Note que este índice no cuantifica la rentabilidad de un particular plan de calificación de primas, pero sí cuantifica la capacidad del plan de calificación para diferenciar los mejores y los peores riesgos. Asumiendo que una aseguradora tiene precios y/o flexibilidad de suscripción, esto conducirá a una mayor rentabilidad.

3.3.2.2 Indicadores de Asociación

- **Tau de Kendall:** Dado (X_1, X_2) y (X_2, Y_2) independientes e idénticamente distribuidas con cópula C ,

$$\tau = \Pr((X_1 - X_2)(Y_1 - Y_2) > 0) - \Pr((X_1 - X_2)(Y_1 - Y_2) < 0)$$

Sin embargo, una forma de estimar τ de una muestra aleatoria de n pares (x_i, y_i) $i = 1, 2, \dots, n$ definimos la siguiente variable indicadora

$$A_{ij} = \text{sgn}(x_i - x_j)(y_i - y_j)$$

se puede notar que

$$\tau = E(A_{ij}) = (+1)Pr((x_i - x_j)(y_i - y_j) > 0) + (-1)Pr((x_i - x_j)(y_i - y_j) < 0)$$

De ello se deduce que un estimador insesgado del coeficiente de Kendall es la llamada τ de Kendall muestral:

$$\frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j>i} A_{ij}$$

- **Rho de Spearman:** Dado (X_1, X_2) , (X_2, Y_2) y (X_3, Y_3) independientes e idénticamente distribuidas con cópula C ,

$$\rho = 3[Pr((X_1 - X_2)(Y_1 - Y_3) > 0) - Pr((X_1 - X_2)(Y_1 - Y_3) < 0)]$$

Sin embargo, una forma de estimar ρ de una muestra aleatoria de n pares (x_i, y_i) $i = 1, 2, \dots, n$ y recordando que ρ_s es la correlación de rango, uno puede cambiar a los rangos de las variables muestrales:

$$R_i = rank(X_i), \quad S_i = rank(Y_i)$$

donde la clasificación se tiene que hacer de orden ascendente. Al hacerlo se puede obtener la siguiente ρ muestral

$$\frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}}$$

Teniendo en cuenta el hecho de que los rangos de n datos son los primeros n números enteros, la expresión anterior se simplifica en

$$1 - 6 \frac{\sum_{i=1}^n (R_i - S_i)^2}{n(n^2 - 1)}$$

es la versión muestral de ρ y es un estimador insesgado.

3.3.2.3 Indicadores de Desempeño

- **Log-Likelihood:** Para cualquier conjunto dado de coeficientes, un GLM implica una media probabilística para cada registro. Eso, junto con el parámetro de dispersión y la forma de distribución elegida, implica una distribución de probabilidad completa. Por lo tanto, es posible calcular, para cualquier registro, la probabilidad (o densidad de probabilidad) que el GLM asignaría a la ocurrencia del resultado real que efectivamente ocurrió. Multiplicar esos valores en todos los registros produciría la probabilidad de que ocurran todos los resultados históricos; este valor se llama *log-likelihood*.

Un GLM se ajusta encontrando el conjunto de parámetros para los cuales la probabili-

dad es la más alta. Esto es intuitivo; en ausencia de otra información, el mejor modelo es el que asigna la mayor probabilidad a los resultados históricos. Dado que la probabilidad suele ser un número extremadamente pequeño, el logaritmo de probabilidad, o *log-likelihood*, suele utilizarse en su lugar para que trabajar con él sea más manejable.

El *log-likelihood* por sí mismo puede ser difícil de interpretar. Por lo tanto, es útil relacionar el *log-likelihood* con sus límites superior e inferior hipotéticos alcanzables con los datos proporcionados.

En el extremo inferior de la escala se encuentra el *log-likelihood* del modelo nulo o **null model**, o un modelo hipotético sin predictores, solo una intersección. Tal modelo produciría la misma predicción para cada registro: la gran media. En el otro extremo se encuentra el modelo saturado, o un modelo hipotético con un número igual de predictores que registros en el conjunto de datos.

- **Devianza:** Esta medida para GLM se define como

$$\text{Devianza} = 1 \times (ll_{\text{saturado}} - ll_{\text{modelo}})$$

donde ll_{saturado} es el *log-likelihood* del modelo saturado, y ll_{modelo} es el *log-likelihood* del modelo que está siendo evaluado. Esto puede expresarse más formalmente de la siguiente manera:

$$D = 2 \times \sum_{i=1}^n \ln f(y_i | \mu_i = y_i) - \ln f(y_i | \mu_i = \mu_i)$$

3.4 Operacionalización y Matriz de consistencia

3.4.1 Cuadro de operacionalización de variables

TABLA N° 3.1: Cuadro de operacionalización de variables

Variables	Definición Conceptual	Dimensiones	Sub dimensiones	Indicadores	Tipo de variable
Robustez	Se dice que una metodología es robusta respecto de las desviaciones de los supuestos del modelo, cuando el proceso continúa trabajando bien, aún cuando, en mayor o menor extensión, los supuestos no se mantienen.	Estabilidad	Precision	Quantile Plots	Razón
			Siniestralidad	Loss ratio Chart	Razón
			Desigualdad de riesgos	Indice de Gini	Razón
Asociación	La asociación estadística mide el grado de dependencia entre dos riesgos o variables.	Dependencia	Correlación por rangos	Tau de Kendall	Razón
Desempeño	Comparación de los modelos candidatos, que se descartando mediante el uso de estadísticos que comparan el poder predictivo de los modelos frente a los datos observados.	Probabilidad	Máxima Verosimilitud	Log-likelihood	Razón
			Cociente MVLE	Devianza	Razón

3.4.2 Matriz de consistencias

TABLA N° 3.2: Matriz de consistencia

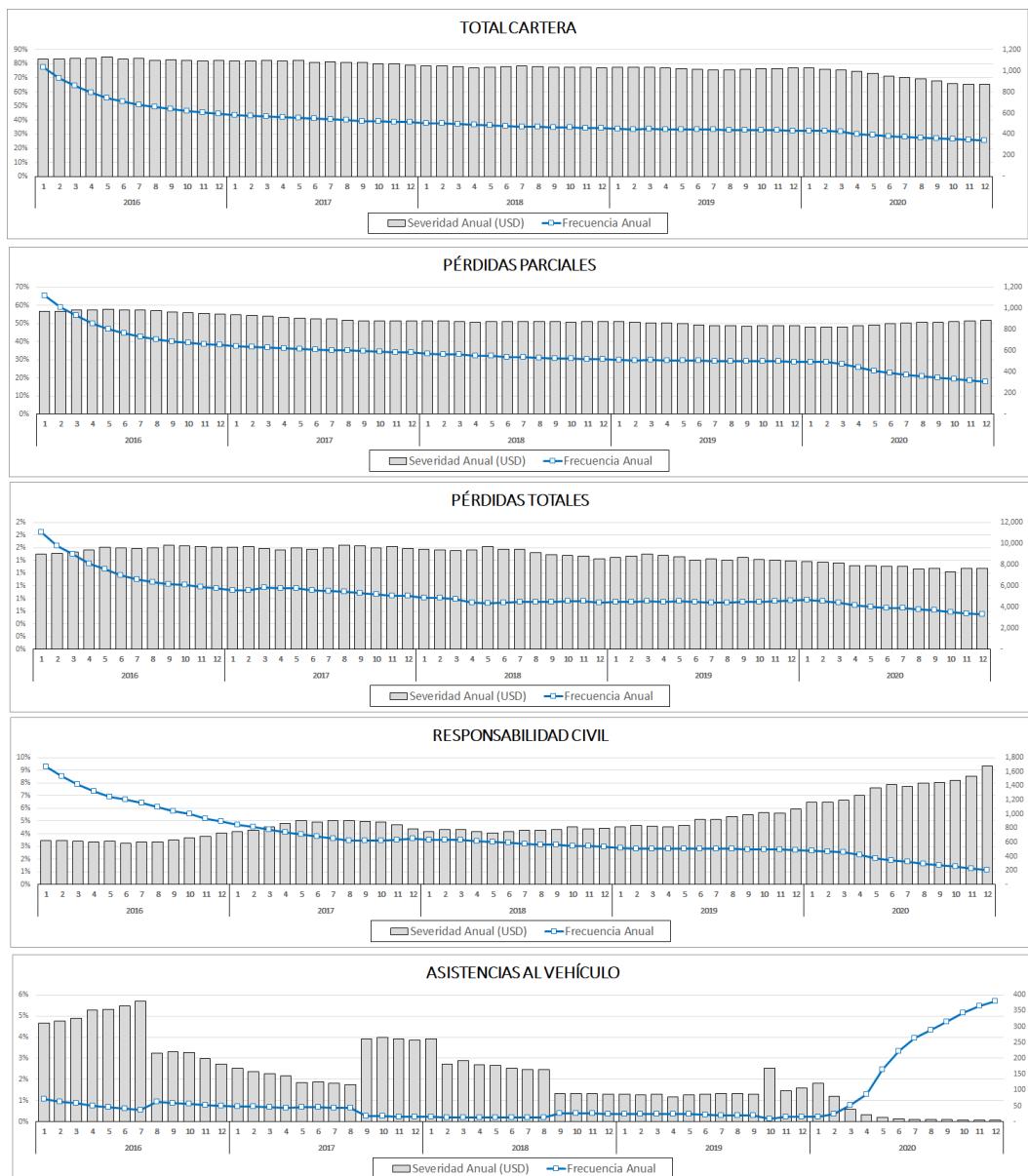
Problemas	Objetivos	Hipótesis
General	General	General
Es posible determinar una propuesta metodológica del cálculo de la prima pura de riesgo sujeto a dependencia entre los riesgos de frecuencia y severidad, y agregados mediante funciones cópulas por cobertura que nos permita estimaciones robustas y de acuerdo con el nivel de riesgo específico de los asegurados frente supuesto usual de independencia	Determinar una propuesta metodológica del cálculo de la prima pura de riesgo sujeto a dependencia entre los riesgos de frecuencia y severidad, y agregados mediante funciones cópulas por cobertura que nos permita estimaciones robustas y de acuerdo con el nivel de riesgo específico de los asegurados frente supuesto usual de independencia.	La propuesta metodológica del cálculo de la prima pura de riesgo sujeto a dependencia entre los riesgos de frecuencia y severidad, y agregados mediante funciones cópulas por cobertura nos permita estimaciones robustas y de acuerdo con el nivel de riesgo específico de los asegurados frente supuesto usual de independencia.
Específicas	Específicas	Específicas
¿Es posible determinar un modelo de regresión conjunto de la frecuencia y de la severidad para estimar la prima pura de riesgo de una determinada cobertura?	Determinar un modelo de regresión conjunto de la frecuencia y de la severidad para estimar la prima pura de riesgo de una determinada cobertura.	Se puede construir un modelo de regresión conjunto de la frecuencia y de la severidad para estimar la prima pura de riesgo de una determinada cobertura.
¿Cómo podemos medir el grado de dependencia entre la frecuencia y la severidad?	Medir el grado de dependencia entre la frecuencia y la severidad.	Se puede determinar el grado de dependencia entre la frecuencia y la severidad.
¿Es posible determinar una propuesta metodológica para el cálculo de la prima total mediante agregación de los riesgos asociados de cada cobertura mediante funciones cópulas?	Determinar una propuesta metodológica de cálculo de la prima total mediante agregación de los riesgos asociados de cada cobertura mediante funciones cópulas.	La primas total se puede obtener mediante agregación de los riesgos asociados a cada cobertura mediante funciones cópulas.

CAPÍTULO IV: ANÁLISIS Y RESULTADOS

4.1 Análisis descriptivo de los datos

Para entender la información procederemos analizar la frecuencia y la severidad por cada una de las coberturas identificadas en la base de datos. El periodo de análisis comprende desde el 01 de enero de 2016 hasta el 31 de diciembre de 2020, de acuerdo a los datos obtenidos de: <https://www2.susep.gov.br/menuestatistica/autoseg/principal.aspx>

FIGURA N° 4.1: Frecuencia y Severidad por cobertura y periodo de análisis



Podemos apreciar que para la Cartera Total la frecuencia está alrededor del 40% en promedio y una severidad 1,000 USD; sin embargo, si lo separamos por cobertura, en el caso de las

Pérdidas Parciales la frecuencia es 30 % en promedio y una severidad 890 USD en promedio, para Pérdidas Parciales la frecuencia es de 0.8 % en promedio y una severidad de 8,600 USD, para Responsabilidad Civil la frecuencia es de 3.1 % en promedio y una severidad de 1,100 USD y por últimos las Asistencias al Vehículo tienen una frecuencia de 1.6 % en promedio y una severidad de 123 USD. Es posible notar un comportamiento distinto en el año 2020 por motivos de la pandemia del COVID-19, ya que siendo una estadística de datos reales es posible este efecto.

Este análisis nos indica que debemos modelar de forma distinta tanto la frecuencia y la severidad por cobertura, además, descartar el año 2020 y realizar una predicción para los años futuros tomando los años desde el 2016 hasta el 2019, ya que por experiencia sabemos que los niveles prepandemia se recuperan.

A continuación procedemos a describir las variables que se incluirán en los análisis y el modelamiento:

Tipo de póliza: Esta variable nos indica la forma de cómo se comercializa la póliza, que puede ser *Individual* o *Colectiva*. Una póliza individual generalmente solo cubre un vehículo, en cambio una póliza colectiva suele cubrir un colectivo de vehículos, generalmente de empresas y/o instituciones.

TABLA N° 4.1: Variable: Tipo de póliza

Cobertura	Niveles	Expuestos	Frecuencia	Severidad ¹	PMPY ²
Pérdida Parciales	Individual	227,892	18%	867	156
	Colectivo	817,699	26%	906	234
Pérdida Totales	Individual	227,892	0.50%	6,579	33
	Colectivo	817,699	0.68%	9,324	63
Responsabilidad Civil	Individual	227,892	1.93 %	1,513	29
	Colectivo	817,699	2.50 %	793	20
Asistencias	Individual	227,892	0.63 %	139	0.9
	Colectivo	817,699	1.61 %	33	0.5

¹En moneda USD

²Per Member Per Year: Monto incurrido de siniestro por cada unidad expuesta

Antiguedad del vehículo: Esta variable nos indica la edad del vehículo desde su fecha de fabricación hasta la fecha de emisión. En la base de datos se hallaron antiguedades de hasta 50 años, sin embargo, para su mejor análisis y rendimiento en el modelo estas son agrupadas, esta conversión se muestra en el anexo A. En la tabla 4.2 se muestran las antiguedades agrupadas.

TABLA N° 4.2: Variable: Antiguedad del vehículo

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	0	41,384	36%	859	305
	1	149,311	34%	896	308
	2	152,047	30%	940	279
	3	135,956	26%	970	251
	4	108,262	24%	926	222
	5	86,081	23%	945	219
	6	66,711	21%	871	185
	7	56,631	20%	840	168
	8	49,161	18%	856	157
	9	39,492	17%	814	138
	10	31,103	16%	782	123
	11-13	55,661	14%	801	115
	14-19	34,936	10%	682	68
	20+	38,854	5%	451	21
Pérdida Totales	0-1	190,695	1%	9443	80
	2	152,047	1%	8820	65
	3-5	330,299	1%	9636	62
	6	66,711	1%	9257	55
	7-16	254,595	1%	7336	37
	17+	51,243	0%	3309	9
Responsabilidad Civil	0-1	190,695	3%	736	19
	2-9	694,341	2%	950	23
	10-13	86,764	2%	997	22
	14+	73,790	2%	1085	20
Asistencias	0	41,384	1%	80	0.54
	1-2	301,358	1%	71	0.80
	3	135,956	1%	58	0.72
	4-5	194,343	1%	41	0.54
	6	66,711	2%	29	0.46
	7	56,631	2%	27	0.48
	8+	249,207	2%	24	0.46

Canales de venta: Esta variable nos indica el medio por el cual se realizó la venta de póliza. Se identificaron cinco canales principales: Banca-Digital, Instituciones, Agentes, Bróker, Exclusivos y otros. En la tabla 4.3 mostramos la frecuencia y la severidad por cada nivel.

TABLA N° 4.3: Variable: Canal de venta

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	Agentes	167,733	27%	903	242
	Banca-Digital	293,233	25%	893	226
	Bróker	290,420	22%	953	210
	Exclusivos	190,518	28%	845	232
	Instituciones	76,402	9%	900	83
	Otros	27,285	33%	879	293
Pérdida Totales	Agentes	167,733	1%	9317	72
	Banca-Digital	293,233	1%	9031	62
	Bróker	290,420	1%	9784	54
	Exclusivos	190,518	1%	7664	49
	Instituciones	76,402	0%	5620	25
	Otros	27,285	1%	9260	84
Responsabilidad Civil	Agentes	167,733	2%	790	19
	Banca-Digital	293,233	2%	675	14
	Bróker	290,420	2%	1058	26
	Exclusivos	190,518	3%	1042	33
	Instituciones	76,402	1%	1099	11
	Otros	27,285	3%	1055	36
Asistencias	Agentes	167,733	1%	44	0.61
	Banca-Digital	293,233	2%	27	0.40
	Bróker	290,420	1%	62	0.85
	Exclusivos	190,518	1%	42	0.59
	Instituciones	76,402	0%	47	0.13
	Otros	27,285	4%	46	1.74

Uso del vehículo: Esta variable nos indica el tipo de uso que se realiza con el vehículo. Se hallaron 20 usos en la base de datos, de las cuales lo reducimos en solo 7 usos. En la Tabla 4.4 mostramos la frecuencia y severidad por cada nivel.

TABLA N° 4.4: Variable: Uso del vehículo

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	Particular	621,227	2 %	24	0.47
	Alquiler	31,250	1 %	109	1.56
	Taxi	43,206	1 %	31	0.35
	Comercial	96,993	1 %	62	0.35
	Transporte Personal	40,394	0 %	124	0.50
	Otros-Usos	51,518	0 %	86	0.34
	Carga	161,002	0 %	520	1.30
Pérdida Totales	Alquiler	31,250	34 %	817	275
	Particular	621,227	29 %	847	248
	Transporte Personal	40,394	23 %	792	182
	Otros-Usos	51,518	22 %	819	177
	Comercial	96,993	15 %	799	119
	Carga	161,002	12 %	736	88
	Taxi	43,206	11 %	1948	223
Responsabilidad Civil	Alquiler	31,250	1 %	7577	91
	Taxi	43,206	1 %	11793	120
	Comercial	96,993	1 %	4582	42
	Particular	621,227	1 %	5898	39
	Transporte Personal	40,394	0 %	11553	48
	Carga	161,002	0 %	12131	40
	Otros-Usos	51,518	0 %	23529	56
Asistencias	Particular	621,227	3 %	599	17
	Carga	161,002	2 %	1447	32
	Otros-Usos	51,518	2 %	717	15
	Comercial	96,993	2 %	957	15
	Transporte Personal	40,394	1 %	1646	22
	Alquiler	31,250	1 %	2392	31
	Taxi	43,206	1 %	1894	23

Clase del vehículo: Esta variable nos indica si el vehículo es Liviano, Pesado o un Vehículo Menor. En la Tabla 4.5 mostramos la frecuencia y severidad por cada nivel.

TABLA N° 4.5: Variable: Clase del vehículo

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	Livianos	765,062	28%	824	231
	Veh. Menor	56,112	6%	499	30
	Pesados	207,983	14%	1547	222
Pérdida Totales	Livianos	765,062	28%	8223	2309
	Veh. Menor	56,112	6%	1798	106
	Pesados	207,983	14%	22369	3217
Responsabilidad Civil	Livianos	765,062	28%	651	183
	Veh. Menor	56,112	6%	965	57
	Pesados	207,983	14%	2133	307
Asistencias	Livianos	765,062	28%	29	8
	Veh. Menor	56,112	6%	36	2
	Pesados	207,983	14%	426	61

Tipo del vehículo: Esta variable nos muestra de una forma más desagregada la característica del vehículo. Se hallaron más de 20 tipos de vehículos los cuales fueron agrupadas de forma distinta por cobertura. En la Tabla 4.6 mostramos la frecuencia y severidad por cada nivel.

TABLA N° 4.6: Variable: Tipo de vehículo

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	Camioneta Rural	297,548	30%	22	7
	Automóvil	325,220	28%	27	8
	Camioneta Pick Up	116,260	23%	32	7
	Otros Livianos	26,033	22%	50	11
	Motocicletas	54,185	6%	47	3
	Otros Veh. Menores	1,926	0%	0	0
	Omnibus	22,649	29%	2	1
	Microbus	21,914	18%	21	4
	Camion	108,037	12%	9	1
Pérdida Totales	Remolcador y Semirremolque	55,384	12%	4	0
	Camioneta Pick Up	116,260	1%	50	0
	Automóvil	325,220	1%	98	1
	Camioneta Rural	297,548	1%	120	1
	Otros Livianos	26,033	1%	259	1
Responsabilidad Civil	Motocicletas	54,185	1%	22	0
	Otros Veh. Menores	1,926	0%	0	0
	Pesados	207,983	0%	44	0
	Camioneta Panel	21,780	3%	1989	60
	Camioneta Rural	297,548	3%	1496	43
	Camioneta STW	4,253	3%	0	0
	Automóvil	325,220	3%	1240	33
Asistencias	Camioneta Pick Up	116,260	2%	845	13
	Motocicletas	54,185	1%	288	2
	Otros Veh. Menores	1,926	0%	0	0
	Omnibus	22,649	3%	612	16
	Remolcador y Semirremolque	55,384	2%	186	4
	Camion	108,037	2%	383	8
	Microbus	21,914	1%	1106	16
	Automóvil	325,220	2%	1031	23
	Camioneta Rural	297,548	2%	1277	23
	Otros Livianos	142,294	1%	831	7
	Veh. Menores	56,112	0%	207	1
	Pesados	207,983	0%	296	1

Marca del Vehículo: Esta variable nos indica la marca del vehículo asegurado. En la base de datos se hallaron diverasas marcas, y por ser tan numerosos, lo que se hizo fue realizar un análisis pareto y centrarnos en las principales marcas, ademas, esto se hizo por clase de vehículo. En la base de datos no se tenía el dato del modelo asociado a la marca. En el anexo A mostraremos el detalle de las marcas utilizadas y modelo utilizado para su agrupación. En la Tabla 4.7 mostramos la frecuencia y severidad por cada nivel.

TABLA N° 4.7: Variable: Marca del vehículo y por clase

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	L1	44,293	41%	1042	424
	L2	72,561	35%	1029	355
	L3	541,482	28%	797	220
	L4	80,629	22%	619	139
	L5	26,097	16%	763	125
	P1	463	54%	343	185
	P2	839	40%	645	261
	P3	37,516	24%	1491	354
	P4	130,559	14%	1634	224
	P5	38,607	6%	1361	87
Pérdida Totales	M1	1,040	28%	1256	354
	M2	4,103	13%	1067	141
	M3	50,968	5%	285	14
	L1	1,679	2%	5076	82
	L2	6,327	1%	7496	83
	L3	271,591	1%	8663	77
	L4	485,465	1%	7889	45
	P1	2,738	1%	17958	164
	P2	13,839	1%	16549	90
	P3	76,503	0%	24093	92
Responsabilidad Civil	P4	114,904	0%	22538	50
	M1	1,450	6%	3303	189
	M2	15,410	2%	1318	28
	M3	39,252	1%	1885	17
	L1	26,132	4%	526	21
	L2	52,559	3%	509	17
	L3	278,741	3%	555	17
	L4	197,972	2%	695	17
	L5	209,659	2%	894	18
	P1	52,383	3%	2382	81
Asistencias	P2	56,270	2%	2050	42
	P3	99,330	2%	1899	29
	M1	2,395	2%	654	13
	M2	6,647	1%	1640	20
	M3	15,839	1%	907	7
	M4	31,231	0%	655	2
	L1	271,436	3%	16	0.40
	L2	281,532	2%	28	0.46
	L3	206,342	1%	64	0.76
	L4	5,753	1%	36	0.26
	P1	1,416	1%	100	0.63
	P2	8,646	0%	206	1.02
	P3	39,455	0%	165	0.65
	P4	119,041	0%	602	1.39
	P5	39,426	0%	281	0.09
	M1	1,040	2%	50	1.15
	M2	10,063	1%	46	0.34
	M3	45,009	0%	23	0.04

Zona demográfica: Esta variable nos indica la zona demográfica donde vive el asegurado. Lo ideal para esta variable sería registrar la zona de circulación del asegurado, pero normalmente no este campo no existe o está incompleto. Todas las zonas halladas se codifican con A para referirse a las provincias y con D a los distritos de mayor densidad. En la Tabla 4.8 mostramos la frecuencia y severidad por cada nivel.

TABLA N° 4.8: Variable: Zona demográfica

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	A1	756	36 %	701	253
	A2	474,830	28 %	789	219
	A3	236,273	23 %	925	213
	A4	222	16 %	840	136
	A5	51,109	23 %	743	174
	A6	10,301	21 %	1028	221
	D1	81,738	22 %	1111	248
	D2	159,887	18 %	1200	219
	D3	10,607	14 %	1320	189
	D4	19,869	11 %	1668	187
Pérdida Totales	A1	6,655	1 %	10749	132
	A2	83,219	1 %	9813	94
	A3	622,206	1 %	8157	48
	A4	51,109	1 %	7385	43
	A5	10,301	1 %	8252	75
	D1	73,410	1 %	7358	63
	D2	73,280	1 %	11081	75
	D3	101,730	1 %	11856	62
	D4	23,680	0 %	13090	46
	D5				
Responsabilidad Civil	A1	587	5 %	817	38
	A2	501,513	3 %	691	22
	A3	209,612	2 %	1367	29
	A4	51,477	3 %	992	30
	A5	10,301	1 %	1961	29
	D1	5,828	2 %	404	10
	D2	61,350	1 %	1533	20
	D3	64,764	1 %	1670	17
	D4	112,674	1 %	1334	11
	D5	27,485	1 %	1492	8
Asistencias	A1	293,398	3 %	15	0.39
	A2	112,298	2 %	21	0.38
	A3	306,385	1 %	76	0.79
	A4	51,109	1 %	26	0.25
	A5	10,301	1 %	72	0.50
	D1	5,828	2 %	22	0.38
	D2	18,112	1 %	103	0.98
	D3	12,527	1 %	108	0.71
	D4	227,769	0 %	277	0.84
	D5	7,865	0 %	431	0.60

Género del asegurado: Esta variable nos indica el género del asegurado, que no necesariamente es el conductor del vehículo. Esta variable tendrá un nivel de "No Aplica", y está relacionado al tipo de persona jurídica, y en algunos casos por falta de información, pero

es de cantidad despreciable. En la Tabla 4.9 mostramos la frecuencia y severidad por cada nivel.

TABLA N° 4.9: Variable: Género del asegurado

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	Femenino	642,185	23 %	932	214
	No Aplica	5,146	20 %	962	190
	Masculino	398,259	26 %	853	221
Pérdida Totales	Femenino	642,185	1 %	10076	56
	No Aplica	5,146	1 %	8257	51
	Masculino	398,259	1 %	7433	57
Responsabilidad Civil	Femenino	642,185	2 %	1023	24
	No Aplica	5,146	1 %	930	7
	Masculino	398,259	2 %	762	19
Asistencias	Femenino	642,185	1 %	48	0.65
	No Aplica	5,146	1 %	70	0.84
	Masculino	398,259	1 %	37	0.55

Tipo de Persona: Esta variable nos indica si el asegurado es una Persona Natural o una Persona Jurídica (Empresas). En la Tabla 4.10 mostramos la frecuencia y severidad por cada nivel.

TABLA N° 4.10: Variable: Tipo de Persona

Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	Empresa	550,779	20 %	1028	208
	Personas	494,811	28 %	798	226
Pérdida Totales	Empresa	550,779	1 %	10467	60
	Personas	494,811	1 %	7387	52
Responsabilidad Civil	Empresa	550,779	2 %	1321	28
	Personas	494,811	3 %	566	15
Asistencias	Empresa	550,779	1 %	77	0.71
	Personas	494,811	2 %	26	0.50

Edad del asegurado: Esta variable nos permitirá captar el comportamiento de la frecuencia de acuerdo a la edad del asegurado. En la Tabla 4.11 mostramos la frecuencia y severidad por cada nivel.

Suma Asegurada: Esta variable se refiere al valor del vehículo al momento de la suscripción. Para su análisis se dividió en rangos sugeridos, con la finalidad de obtener un comportamiento uniforme por cada cobertura. Estos rangos son mostrados en el Anexo A, y se puede observar como la frecuencia se comporta distinto por cada cobertura, por ejemplo, para las Pérdidas Parciales la relación directa, sin embargo, para el resto de coberturas la relación es inversa.

TABLA N° 4.11: Variable: Edad del asegurado

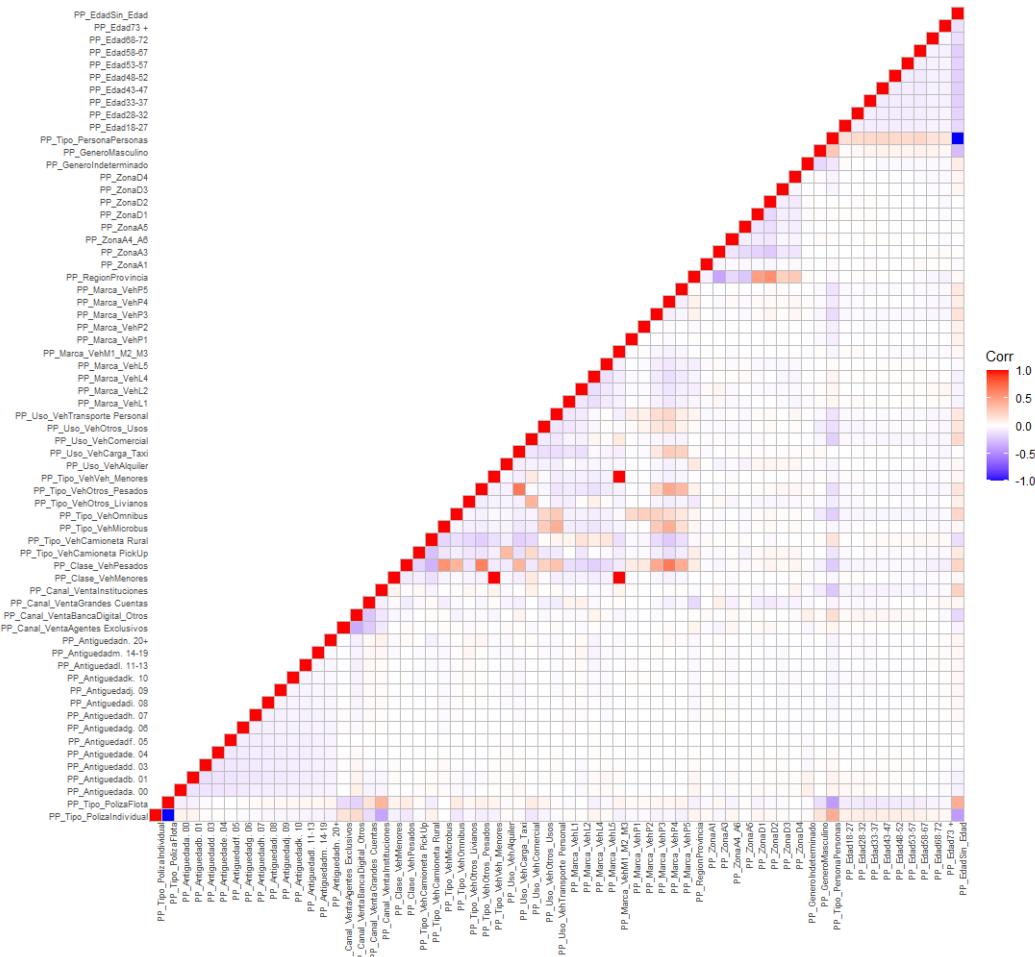
Cobertura	Niveles	Expuestos	Frecuencia	Severidad	PMPY
Pérdida Parciales	18-27	29,717	36%	838	304
	28-32	57,175	34%	797	269
	33-37	67,212	32%	804	261
	38-42	69,025	30%	805	240
	43-47	62,269	28%	800	227
	48-52	52,422	27%	799	213
	53-57	45,667	26%	767	200
	58-67	68,686	24%	778	184
	68-72	20,467	21%	818	169
	73 +	22,172	18%	764	135
Sin Edad		550,779	20%	1028	208
Pérdida Totales	18-27	29,717	1%	6808	92
	28-32	57,175	1%	6972	75
	33-37	67,212	1%	7371	61
	38-42	69,025	1%	8392	61
	43-52	114,691	1%	7439	45
	53-57	45,667	1%	7451	40
	58-62	38,181	0%	8382	39
	63-77	62,800	0%	6828	30
	78 +	10,343	0%	4644	14
	Sin Edad	550,779	1%	10467	60
Responsabilidad Civil	18-27	29,717	3%	461	16
	28-32	57,175	3%	610	17
	33-37	67,212	3%	553	15
	38-42	69,025	3%	626	16
	43-52	114,691	3%	576	15
	53-57	45,667	3%	542	14
	58-62	38,181	3%	551	15
	63-67	30,505	2%	502	12
	68-72	20,467	3%	507	13
	73-77	11,828	3%	861	22
Asistencias	78-82	6,239	3%	424	11
	83-87	2,669	2%	338	8
	88+	1,435	3%	445	12
	Sin Edad	550,779	2%	1321	28
	18-27	29,717	2%	55	0.89
	28-32	57,175	2%	44	0.75
	33-37	67,212	2%	34	0.60
	38-42	69,025	2%	29	0.51
	43-47	62,269	2%	21	0.40
	48-52	52,422	2%	23	0.45
	53-57	45,667	2%	25	0.49
	58-67	68,686	2%	13	0.28
	68-72	20,467	2%	14	0.33
	73-82	18,067	3%	6	0.17
	83+	4,104	3%	9	0.32
Sin Edad		550,779	1%	77	0.71

4.2 Análisis de los datos

4.2.1 Modelos Marginales de la Frecuencia y de la Severidad

Matriz de correlaciones De las variables analizadas procedemos a analizar la matriz de correlación para verificar posible multicolinealidad entre las variables, por ejemplo, para la cobertura Pérdida Parcial

FIGURA N° 4.2: Matriz de correlación - Pérdida Parcial



la matriz de correlación nos indica que existen variables correlacionadas, de forma positiva y negativa. Esta información se tomará en cuenta para construir el modelo marginal y evitar problemas de multicolinealidad. Aunque, en la codificación se utilizó el paquete *Alias* del R para detectar los problemas de multicolinealidad, sin embargo, las conclusiones fueron similares. Esto se realizó para el resto de coberturas, cuyos resultados se muestran en el Anexo B.

Modelos Lineales Generalizados Marginales Habiendo analizado las posibles causas de multicolinealidad procedemos a construir los modelos marginales GLM de la Frecuencia y de la Severidad por cobertura. Los códigos en R se muestran en el Anexo B.

TABLA N° 4.12: Modelo Marginal de Frecuencia y Severidad - Pérdidas Parciales

var-level	Frecuencia (β)	Severidad (α)	var-level	Frecuencia (β)	Severidad (α)
(Intercept)	-1.4958	6.8387	Marca-VehL3	0.0000	0.0000
Antiguedada. 00	0.1930	-0.0258	Marca-VehL4	-0.1444	-0.1167
Antiguedadb. 01	0.1599	-0.0229	Marca-VehL5	-0.5007	0.1557
Antiguedadc. 02	0.0000	0.0000	Marca-VehM1-M2-M3	-0.6576	0.4385
Antiguedadd. 03	-0.1428	0.0377	Marca-VehP1	0.4069	-0.5971
Antiguedade. 04	-0.2258	-0.0109	Marca-VehP2	0.5295	0.0212
Antiguedadf. 05	-0.2515	0.0250	Marca-VehP3	0.4690	0.3193
Antiguedadg. 06	-0.3098	-0.0067	Marca-VehP4	0.0786	0.2637
Antiguedadh. 07	-0.3498	-0.0193	Marca-VehP5	-0.6642	0.2355
Antiguedadi. 08	-0.3780	0.0094	SAa.[1000-2000>	-1.2163	-0.7024
Antiguedadj. 09	-0.4083	-0.0028	SAb.[2000-3000>	-0.5825	-0.6936
Antiguedadk. 10	-0.4512	-0.0422	SAc.[3000-4000>	-0.4271	-0.5761
Antiguedadl. 11-13	-0.4880	0.0043	SAd.[4000-5000>	-0.3653	-0.4428
Antiguedadm. 14-19	-0.7549	-0.0800	SAe.[5000-6000>	-0.2691	-0.3932
Antiguedadn. 20+	-1.1901	-0.2489	SAf.[6000-7000>	-0.2375	-0.3658
Canal-VentaAgentes	0.0814	0.0102	SAg.[7000-8000>	-0.2229	-0.2953
Canal-VentaBancaDigital-Otros	-0.0701	-0.0047	SAh.[8000-9000>	-0.1767	-0.2781
Canal-VentaBroker	0.0000	0.0000	SAi.[9000-13000>	-0.1302	-0.2048
Canal-VentaExclusivos	0.0989	-0.0127	SAj.[13000-999999]	0.0000	0.0000
Canal-VentaInstituciones	-0.5083	0.0984	Tipo-PolizaFlota	0.1442	-0.1310
Edad18-27	0.1404	0.0531	Tipo-PolizaIndividual	0.0000	0.0000
Edad28-32	0.1090	-0.0190	Uso-VehAlquiler	0.0664	0.0340
Edad33-37	0.0717	-0.0123	Uso-VehCarga-Taxi	-0.8514	0.6021
Edad38-42	0.0000	0.0000	Uso-VehComercial	-0.3486	0.1093
Edad43-47	-0.0250	-0.0335	Uso-VehOtros-Usos	-0.2774	-0.1705
Edad48-52	-0.0749	-0.0187	Uso-VehParticular	0.0000	0.0000
Edad53-57	-0.0962	-0.0573	Uso-VehTransporte Personal	-0.2954	-0.2249
Edad58-67	-0.1921	-0.0366	ZonaA1	0.2952	-0.0801
Edad68-72	-0.2939	0.0246	ZonaA2	0.0000	0.0000
Edad73 +	-0.3475	0.0082	ZonaA3	-0.0485	0.0556
EdadSin-Edad	-0.0412	0.1208	ZonaA4-A6	-0.2146	0.1521
GeneroFemenino	0.0000	0.0000	ZonaA5	0.1168	-0.0924
GeneroIndeterminado	0.1069	0.0573	ZonaD1	-0.2794	0.2113
GeneroMasculino	0.0174	0.0290	ZonaD2	-0.3243	0.2174
Marca-VehL1	0.2538	0.2192	ZonaD3	-0.6388	0.2635
Marca-VehL2	0.1680	0.1789	ZonaD4	-0.5521	0.6066

Los modelos de regresión marginal construidos son la base de nuestro estudio, es por ello que debemos revisar la robustez de los modelos mediante el índice de GINI, estas se muestran en la figura Figura N° 4.3. Como se puede observar, el modelo GLM de la cobertura Pérdida Parcial tiene una capacidad de segmentar a lo malos y buenos riesgos en un 49.11 %, el modelo GLM de la cobertura Pérdida Total tiene una capacidad de 46.82 %, el modelo GLM de la cobertura Responsabilidad Civil tiene una capacidad de 53.28 % y el modelo GLM de la cobertura Asistencias tiene una capacidad de 64.26 %. De acuerdo a la literatura estándar, los valores de los índices obtenidos indican una buena capacidad de poder segmentar los datos.

4.2.2 Pruebas de indicios de Dependencia

En esta sección, mostraremos la existencia de cierta dependencia entre la frecuencia y la severidad, que usualmente suelen asumirse como independientes. Utilizaremos en primera instancia gráficos para revelar la dependencia.

TABLA N° 4.13: Modelo Marginal de Frecuencia y Severidad - Pérdidas Totales

var-level	Frecuencia (β)	Severidad (α)	var-level	Frecuencia (β)	Severidad (α)
(Intercept)	-6.0207	8.8907	Marca-VehM3	0.6250	-1.1767
Edad18-27	0.6982	-0.0300	Marca-VehP1-P2-P3-P4	-0.7029	0.6995
Edad28-32	0.4720	-0.0070	Tipo-PolizaFlota	-0.4449	-0.0767
Edad33-42	0.1939	0.0781	Tipo-PolizaIndividual	0.0000	0.0000
Edad43-52	0.0000	0.0000	Uso-VehAlquiler	0.5607	0.3007
Edad53-77	-0.2055	0.0419	Uso-VehAmbul-TransIntProv-Carga-Tur-Esc	0.3141	0.2627
Edad78 +	-0.7924	-0.1815	Uso-VehComercial-Taxi-TransPers	0.1760	-0.1017
EdadSin-Edad	0.1872	0.2343	Uso-VehOtros-Usos	-0.7710	-0.8111
GeneroFemenino	0.0000	0.0000	Uso-VehParticular	0.0000	0.0000
GeneroMasculino	0.1818	0.0323	ZonaA1	0.7206	0.3403
Marca-VehL1	0.8985	-0.3399	ZonaA2-A5	0.4227	0.0544
Marca-VehL2	0.6156	0.0507	ZonaA3	0.0000	0.0000
Marca-VehL3	0.3973	0.1151	ZonaA4	0.1341	-0.1399
Marca-VehL4	0.0000	0.0000	ZonaD1	0.1883	0.1120
Marca-VehM1	2.1597	-0.5939	ZonaD2-D3	-0.1291	0.0208
Marca-VehM2	1.4703	-1.4691	ZonaD4	-0.5374	0.3940

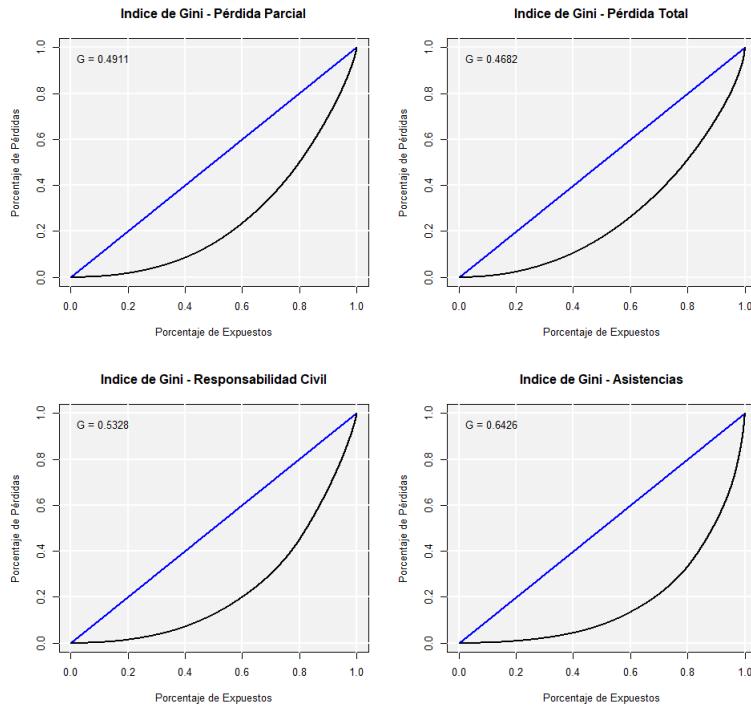
TABLA N° 4.14: Modelo Marginal de Frecuencia y Severidad - Responsabilidad Civil

var-level	Frecuencia (β)	Severidad (α)	var-level	Frecuencia (β)	Severidad (α)
(Intercept)	-4.5046	6.8322	Marca-VehP1-P2	0.3242	0.6791
Antiguedada. 0-1	0.1522	0.0148	Marca-VehP3	-0.1802	0.4541
Antiguedadb. 2-9	0.0000	0.0000	Uso-VehSerenazgo	0.8643	0.2877
Antiguedadc. 10+	-0.1068	-0.2256	Uso-VehTransInterProv-Urb	0.1963	-0.1287
Canal-VentaOtras	0.2161	-0.2133	Uso-VehParticular	0.0000	-2.2461
Canal-VentaExclusivos-Agentes	0.0435	0.0000	Uso-VehAlq-Amb-Carg-Com	-0.2263	0.0000
Canal-VentaBroker	0.0000	0.0213	Uso-VehTransporte Personal	-0.5190	0.2859
Canal-VentaBanca y Digital	-0.0923	0.2788	Uso-VehTurismo	-0.5698	0.0466
Canal-VentaInstituciones	-0.5771	-0.0148	Uso-VehEscolar	-0.5963	0.6057
Edad18-42	0.0740	0.0657	Uso-VehTaxi	-0.6426	0.4276
Edad43-52	0.0000	0.0000	Uso-VehInstrucción	-3.0233	0.0415
Edad53+	-0.0710	0.0962	ZonaA1	0.3943	0.6509
EdadSin-Edad	0.0108	0.2366	ZonaA2	0.0000	0.0000
Marca-VehL1	0.2404	-0.0516	ZonaA3	-0.2704	0.3174
Marca-VehL2	0.0879	-0.0326	ZonaA4	0.0804	0.1037
Marca-VehL3	0.0000	0.0000	ZonaA5	-0.6248	0.8208
Marca-VehL4	-0.0969	0.0594	ZonaD1	-0.3156	-0.2785
Marca-VehL5	-0.1261	0.1812	ZonaD2	-0.7704	0.4737
Marca-VehM1-M2	-0.5064	0.0299	ZonaD3	-0.9557	0.6803
Marca-VehM3	-1.2695	0.5495	ZonaD4	-1.1184	0.4885
Marca-VehM4	-1.6448	-0.0397	ZonaD5	-1.6368	0.5339

TABLA N° 4.15: Modelo Marginal de Frecuencia y Severidad - Asistencias

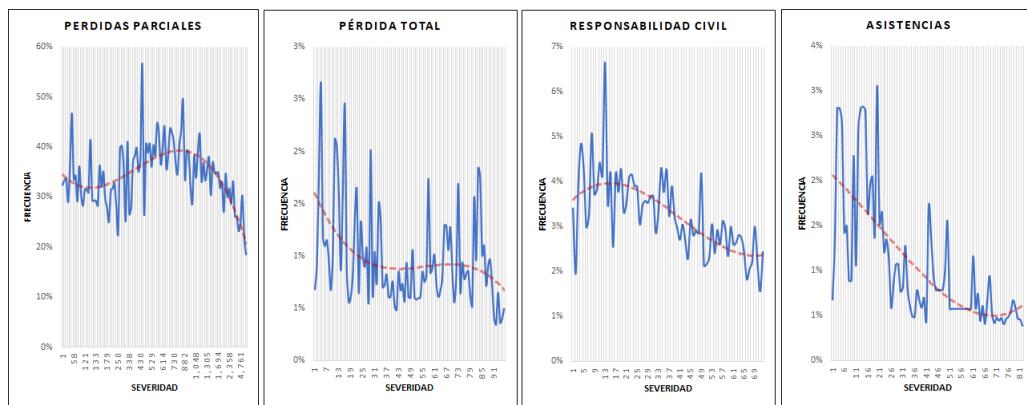
var-level	Frecuencia (β)	Severidad (α)	var-level	Frecuencia (β)	Severidad (α)
(Intercept)	-4.2874	4.6876	Marca-VehP3-P4	-0.6706	1.3200
Antiguedada. 0	-0.3732	0.6929	SAA.[1000-10000>	0.0000	0.0000
Antiguedadb. 1-2	0.0000	0.0000	SAB.[10000-18000>	-0.3577	0.2886
Antiguedadc. 3+	0.0508	-0.0947	SAC.[18000-99999]	-0.5868	0.5299
Canal-VentaAgentes-BancaDigital	0.1045	-0.4080	Uso-VehAlquiler-Escolar	0.6519	0.4431
Canal-VentaBroker	0.0000	0.0000	Uso-VehParticular	0.0000	0.0000
Canal-VentaExclusivos	-0.1341	-0.2050	Uso-VehSerenazgo	1.3647	0.9045
Canal-VentaInstituciones	-1.0202	0.0163	Uso-VehTaxi-Amb-Tur-Com	-0.1338	0.4297
Canal-VentaOtras	0.3212	-0.3016	Uso-VehTransPer-Carga-TransInt-TransUrb-Instr	-0.3862	0.1028
Marca-VehL1	0.0274	-0.2591	ZonaA1-A2	0.1095	-0.3183
Marca-VehL2	0.0000	0.0000	ZonaA3	0.0000	0.0000
Marca-VehL3-L4	-0.2537	0.0225	ZonaA4-A5	-0.1800	0.0915
Marca-VehM1-M2-M3	-1.0790	0.4842	ZonaD1-D4-D5	-0.8697	0.7099
Marca-VehP1	0.3766	-0.0327	ZonaD2-D3	0.2768	0.4620
Marca-VehP2-P5	-1.1444	2.0106			

FIGURA N° 4.3: Robustez de los modelos de GLM Marginales - Índice de GINI



En la Figura N° 4.4 se muestra un primer indicio de una forma de asociación entre la frecuencia y la severidad en valores brutos de los mismos.

FIGURA N° 4.4: Indicio de dependencia entre la Frecuencia y la Severidad por Cobertura



Es posible observar la existencia de una relación u asociación no lineal entre la Frecuencia y la Severidad, por lo que para medir el grado de asociación será medido con la tau de kendal.

Otros gráficos que podemos utilizar para determinar la existencia de dependencia entre dos riesgos son el Chi-Plot y el K-Plot, para los cuales se definieron criterios de dependencia entre dos riesgos:

FIGURA N° 4.5: Chi-Plots de la Frecuencia y Severidad por cobertura

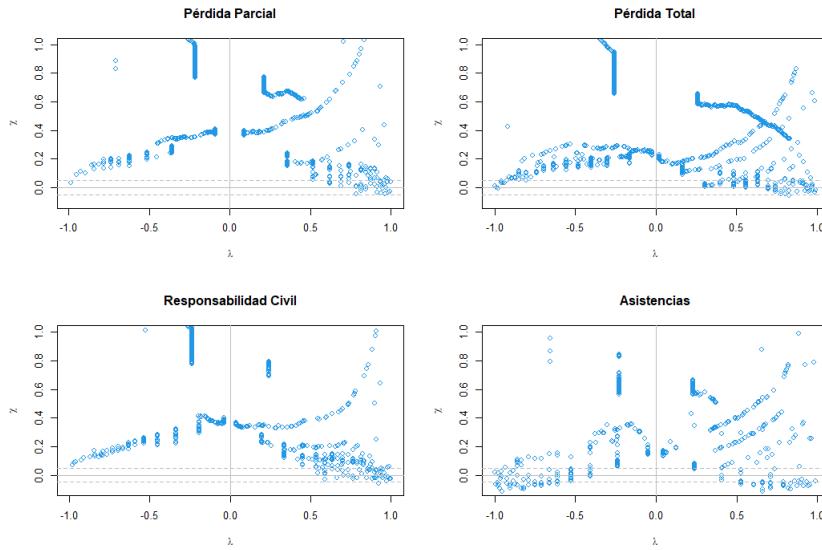
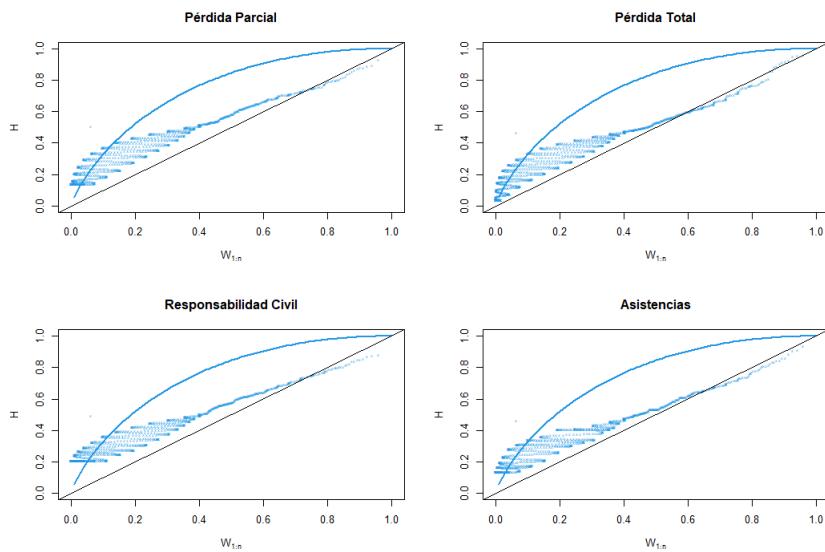


FIGURA N° 4.6: K-Plots de la Frecuencia y Severidad por cobertura



En la Figura N° 4.5 podemos observar que la mayoría de los puntos se encuentran fuera de los rangos al 95% de confianza. De acuerdo a ([Genest y Boies, 2003](#)) si tenemos puntos entrelazados, dentro y fuera de la región de confianza, tenemos una relación de dependencia compleja, lo que implica que utilizar relaciones del tipo lineal serán inservibles. En la Figura N° 4.6 el criterio era que si todos los puntos caían sobre la curva, entonces tendríamos un dependencia perfecta, y sobre la diagonal indicaría riesgos independientes, por lo que el mensaje de estos gráficos es similar, existe cierto nivel de dependencia entre los riesgos. El K-Plot nos da una información adicional, en el sentido que el grado de dependencia es menor en las colas de la distribución.

Con este apartado se concluye la existencia de cierto grado de dependencia entre la frecuencia y la severidad, en particular para el presente trabajo, el tipo de asociación es no lineal, que se hace complejo dado que varía a lo largo del cuerpo de la distribución.

4.2.3 Modelos GLM Conjuntos de la Frecuencia y de la Severidad

Para la construcción del modelo conjunto, utilizaremos como input los modelos GLM marginales de la Frecuencia y de la Severidad. Adicionalmente, calcularemos los valores iniciales de los parámetros para que estos sean optimizados mediante el método de *Log-likelihood Maximum Verosimil*. Entonces por cobertura tenemos:

TABLA N° 4.16: Pérdida Parcial | Log-likelihood de los modelos de regresión

Modelo	Loglikelihood	$\hat{\theta}$	τ	PPR Base
Cópula Gauss	-5,775,746	0.083700	0.053333	403
Cópula Clayton	-5,774,749	0.170904	0.078725	639
Cópula Gumbel	-5,772,251	1.010034	0.009934	509
Cópula Frank	-5,771,951	-0.285000	-0.031616	564

TABLA N° 4.17: Pérdida Total | Log-likelihood de los modelos de regresión

Modelo	Loglikelihood	$\hat{\theta}$	τ	PPR Base
Cópula Gauss	-188,829	0.08368066	0.05333513	92
Cópula Clayton	-196,067	10.48703	0.8398338	20
Cópula Gumbel	-188,835	1.00000	0.00000	95
Cópula Frank	-198,829	0.6069313	0.06692831	89

TABLA N° 4.18: Responsabilidad Civil | Log-likelihood de los modelos de regresión

Modelo	Loglikelihood	$\hat{\theta}$	τ	PPR Base
Cópula Gauss	-767,378	0.063070	0.040179	56
Cópula Clayton	-767,350	0.118702	0.056026	63
Cópula Gumbel	-786,157	14.216610	0.929660	7
Cópula Frank	-768,980	0.856516	0.094193	14

TABLA N° 4.19: Asistencias | Log-likelihood de los modelos de regresión

Modelo	Loglikelihood	$\hat{\theta}$	τ	PPR Base
Cópula Gauss	-288,939	0.07255326	0.04622946	7
Cópula Clayton	-289,311	0.02560843	0.01264234	6
Cópula Gumbel	-288,945	1.00012900	0.00012879	7
Cópula Frank	-288,942	0.36887960	0.04092350	7

En cada una de las tablas se muestran los valores de los *loglikelihood* por cobertura y por tipo de cópula utilizada. bajo el criterio de *loglikelihood* elegimos una familia de cópulas por cobertura. Para las Pérdidas Parciales el mayor *loglikelihood* es de -5,771,951 por lo que elegimos la Cúpula de Frank, para la cobertura Pérdida Total el mayor *loglikelihood* es de -188,829 por lo que elegimos la Cúpula de Gauss, para la cobertura Responsabilidad Civil el mayor *loglikelihood* es de -767,350 por lo que elegimos la Cúpula de Clayton y para

la cobertura de Asistencias el mayor *loglikelihood* es de -288,939 por lo que elegimos la Cúpula de Gauss.

Los códigos en R que se utilizaron para calcular cada modelo de regresión se hallan en detalle en el Anexo C.

Por lo tanto, los modelos de regresión conjunta de la Frecuencia y severidad por cada cobertura y con la familia de copula elegida son las siguientes:

TABLA N° 4.20: Modelo de Regresión GLM Conjunto - Pérdidas Parciales

Cúpula FRANK $\theta = 1.01$			$Loglikelihood = -5,772,251$			$\tau = 0.009934$
var-level	Frecuencia (β)	Severidad (α)	var-level	Frecuencia (β)	Severidad (α)	
(Intercept)	-1.5804	7.9157	Marca-VehL3	0.0000	0.0000	
Antiguedada. 00	0.4223	-0.2577	Marca-VehL4	0.1943	-0.5419	
Antiguedadb. 01	0.1898	0.2016	Marca-VehL5	0.3447	0.2673	
Antiguedad. 02	0.0000	0.0000	Marca-VehM1-M2-M3	-0.0339	0.2058	
Antiguedadd. 03	-0.0424	0.1527	Marca-VehP1	0.4477	-0.6627	
Antiguedade. 04	-0.0915	-0.0761	Marca-VehP2	0.5961	-0.1700	
Antiguedadf. 05	-0.0821	-0.0566	Marca-VehP3	0.3735	0.5000	
Antiguedadg. 06	-0.1114	-0.3866	Marca-VehP4	0.1105	0.0969	
Antiguedadh. 07	-0.0047	-0.3385	Marca-VehP5	-0.0529	-0.0282	
Antiguedad. 08	-0.0267	-0.2672	SAA.[1000-2000>	-0.9033	-0.8563	
Antiguedadj. 09	0.0031	-0.2118	SAB.[2000-3000>	-0.3307	-0.7701	
Antiguedadk. 10	0.2576	-0.1106	SAC.[3000-4000>	-0.0984	-0.6664	
Antiguedadl. 11-13	-0.0913	0.0072	SAD.[4000-5000>	0.0028	-0.5877	
Antiguedadm. 14-19	-0.0318	-0.3697	SAE.[5000-6000>	0.3724	-0.7063	
Antiguedadn. 20+	-0.5157	-0.4655	SAF.[6000-7000>	0.4289	-0.8169	
Canal-VentaAgentes Exclusivos	0.1619	-0.1176	SAG.[7000-8000>	0.4228	-0.8143	
Canal-VentaBancaDigital-Otros	-0.1518	0.2469	SAH.[8000-9000>	0.3204	-1.0248	
Canal-VentaCorredores	0.0000	0.0000	SAI.[9000-13000>	0.0946	-0.5696	
Canal-VentaGrandes Cuentas	0.0627	0.2721	SAJ.[13000-999999]	0.0000	0.0000	
Canal-VentaInstituciones	-0.6424	0.0105				
Edad18-27	0.2530	-0.1066	Tipo-PolizaFlota	0.1055	0.1478	
Edad28-32	0.1611	-0.0151	Tipo-PolizaIndividual	0.0000	0.0000	
Edad33-37	0.1174	0.0211	Uso-VehAlquiler	0.0173	-0.0529	
Edad38-42	0.0000	0.0000	Uso-VehCarga-Taxi	-0.6184	0.3209	
Edad43-47	0.0664	-0.1118	Uso-VehComercial	-0.1140	-0.1903	
Edad48-52	0.0502	-0.2003	Uso-VehOtros-Usos	0.5729	-0.4462	
Edad53-57	0.0749	-0.2402	Uso-VehParticular	0.0000	0.0000	
Edad58-67	-0.0846	-0.1625	Uso-VehTransporte Personal	0.2981	-0.6437	
Edad68-72	0.1578	-0.3279	ZonaA1	0.4260	-0.0549	
Edad73 +	0.0937	-0.3081	ZonaA2	0.0000	0.0000	
EdadSin-Edad	-0.1640	0.3475	ZonaA3	0.1675	-0.2783	
GeneroFemenino	0.0000	0.0000	ZonaA4-A6	0.5853	-0.2157	
GeneroIndeterminado	0.3619	-0.0252	ZonaA5	0.4257	-0.5504	
GeneroMasculino	0.0583	0.0266	ZonaD1	-0.0066	-0.1416	
Marca-VehL1	0.5418	-0.1716	ZonaD2	-0.0783	-0.1510	
Marca-VehL2	0.3597	-0.1209	ZonaD3	-0.0717	-0.0285	
			ZonaD4	-0.0424	0.7375	

4.2.4 Agregación de la Prima Pura de Riesgo por Cobertura

A partir de esta subsección básicamente describiremos la propuesta metodológica para el cálculo de una única Prima Pura de Riesgo en función de los modelos conjuntos de Frecuencia y Severidad, incluyendo su grado de dependencia, calculados en los puntos anteriores.

TABLA N° 4.21: Modelo de Regresión GLM Conjunto - Pérdidas Totales

Cópula GAUSS $\theta = 0.08368066$		$Loglikelihood = -188,829$	$\tau = 0.05333513$		
var-level	Frecuencia (β)	Severidad (α)	var-level	Frecuencia (β)	Severidad (α)
(Intercept)	-5.7924	10.2933	Marca-VehM3	0.7406	-1.8000
Edad18-27	0.7704	-0.2614	Marca-VehP1-P2-P3-P4	-0.7991	0.3155
Edad28-32	0.4645	-0.0641	Tipo-PolizaFlota	-0.4412	-0.2388
Edad33-42	0.1140	0.2567	Tipo-PolizaIndividual	0.0000	0.0000
Edad43-52	0.0000	0.0000	Uso-VehAlquiler	0.6932	-0.3058
Edad53-77	-0.2211	-0.0339	Uso-VehAmbul-TransIntProv-Carga-Tur-Esc	0.6157	-0.0314
Edad78 +	-0.1549	-1.4905	Uso-VehComercial-Taxi-TransPers	0.3858	-0.6809
EdadSin-Edad	-0.0139	1.0199	Uso-VehOtras-Uso	0.8861	-1.8950
GeneroFemenino	0.0000	0.0000	Uso-VehParticular	0.0000	0.0000
GeneroMasculino	0.1247	0.1836	ZonaA1	1.8261	-1.1696
Marca-VehL1	2.9547	-1.7092	ZonaA2-A5	0.5650	-0.6182
Marca-VehL2	2.1613	-0.7824	ZonaA3	0.0000	0.0000
Marca-VehL3	0.4244	-0.0632	ZonaA4	0.7047	-1.2507
Marca-VehL4	0.0000	0.0000	ZonaD1	0.4840	-0.7387
Marca-VehM1	2.8090	-1.5017	ZonaD2-D3	0.0136	-0.5893
Marca-VehM2	1.7898	-2.0125	ZonaD4	0.4061	-1.1410

TABLA N° 4.22: Modelo de Regresión GLM Conjunto - Responsabilidad Civil

Cópula CLAYTON $\theta = 0.1187016$		$Loglikelihood = -767,350$	$\tau = 0.05602563$		
var-level	Frecuencia (β)	Severidad (α)	var-level	Frecuencia (β)	Severidad (α)
(Intercept)	-4.2504	8.3939	Marca-VehP1-P2	0.1291	1.0761
Antiguedada. 0-1	0.3360	-0.5056	Marca-VehP3	-0.3123	0.3273
Antiguedadb. 2-9	0.0000	0.0000	Uso-VehSerenazgo	-0.0826	0.0804
Antiguedadc. 10+	0.0148	-0.7176	Uso-VehTransInterProv-Urb	1.7016	-1.4727
Canal-VentaOtras	-0.0591	-0.1849	Uso-VehParticular	3.1605	-2.5760
Canal-VentaGrandesCuentas-AgentesExc	0.0000	0.0000	Uso-VehAlq-Amb-Carg-Com	0.0000	0.0000
Canal-VentaCorredores	0.0302	0.1189	Uso-VehTransporte Personal	1.3405	-0.2897
Canal-VentaBanca y Digital	-0.2584	-0.0151	Uso-VehTurismo	-0.2753	-0.7953
Canal-VentaInstituciones	0.4135	-0.4016	Uso-VehEscolar	0.6063	0.2579
Edad18-42	-0.0519	0.3687	Uso-VehTaxi	0.4525	-0.5811
Edad43-52	0.0000	0.0000	Uso-VehInstrucción	1.2694	-1.4410
Edad53+	-0.1297	0.1406	ZonaA1	3.2422	-0.5422
EdadSin-Edad	-0.1315	0.4803	ZonaA2	0.0000	0.0000
Marca-VehL1	0.4185	-0.9047	ZonaA3	-0.1559	-0.4263
Marca-VehL2	0.2019	-0.7040	ZonaA4	0.3797	-0.8058
Marca-VehL3	0.0000	0.0000	ZonaA5	0.9758	-0.5989
Marca-VehL4	-0.0762	-0.2283	ZonaD1	1.4755	-1.5941
Marca-VehL5	-0.0950	-0.2114	ZonaD2	-0.2116	-0.7341
Marca-VehM1-M2	0.9316	-1.0202	ZonaD3	-0.4034	-0.5239
Marca-VehM3	-0.7572	-0.7411	ZonaD4	-0.6233	-0.6699
Marca-VehM4	-1.3062	-1.1169	ZonaD5	0.1940	-0.8339

TABLA N° 4.23: Modelo de Regresión GLM Conjunto - Asistencias

Cópula GAUSS $\theta = 0.07255326$		$Loglikelihood = -288,939$	$\tau = 0.04622946$		
var-level	Frecuencia (β)	Severidad (α)	var-level	Frecuencia (β)	Severidad (α)
(Intercept)	-4.1665	6.1156	Marca-VehP3-P4	-1.7171	1.5925
Antiguedada. 0	0.3100	-0.2158	SAa.[1000-10000>	0.0000	0.0000
Antiguedadb. 1-2	0.0000	0.0000	SAb.[10000-18000>	-0.4015	0.0615
Antiguedadc. 3+	-0.0725	0.1895	SAc.[18000-99999]	-0.6796	0.3678
Canal-VentaAgentesExc-BancaDigital	0.0206	-0.2522	Uso-VehAlquiler-Escolar	0.9677	-0.0237
Canal-VentaCorredores	0.0000	0.0000	Uso-VehParticular	0.0000	0.0000
Canal-VentaGrandes Cuentas	-0.1012	-0.4350	Uso-VehSerenazgo	2.2115	0.6802
Canal-VentaInstituciones	-0.2116	-0.8539	Uso-VehTaxi-Amb-Tur-Com	0.3666	-0.4822
Canal-VentaOtras	0.7192	-0.2436	Uso-VehTransPer-Carga-TransInt-TransUrb-Instr	1.1631	-0.9644
Marca-VehL1	0.0646	-0.0913	ZonaA1-A2	0.1550	-0.1266
Marca-VehL2	0.0000	0.0000	ZonaA3	0.0000	0.0000
Marca-VehL3-L4	-0.1085	-0.2256	ZonaA4-A5	0.2712	-1.1097
Marca-VehM1-M2-M3	-0.4148	-0.6801	ZonaD1-D4-D5	-0.7206	0.4472
Marca-VehP1	0.7984	-0.2438	ZonaD2-D3	0.8993	-0.1378
Marca-VehP2-P5	-1.3206	1.4294			

Nuestro primer paso será calcular la Prima Pura de Riesgo por vehículo asegurado utilizando la fórmula que describimos en el teorema 10 de la subsección 2.2.4.3.

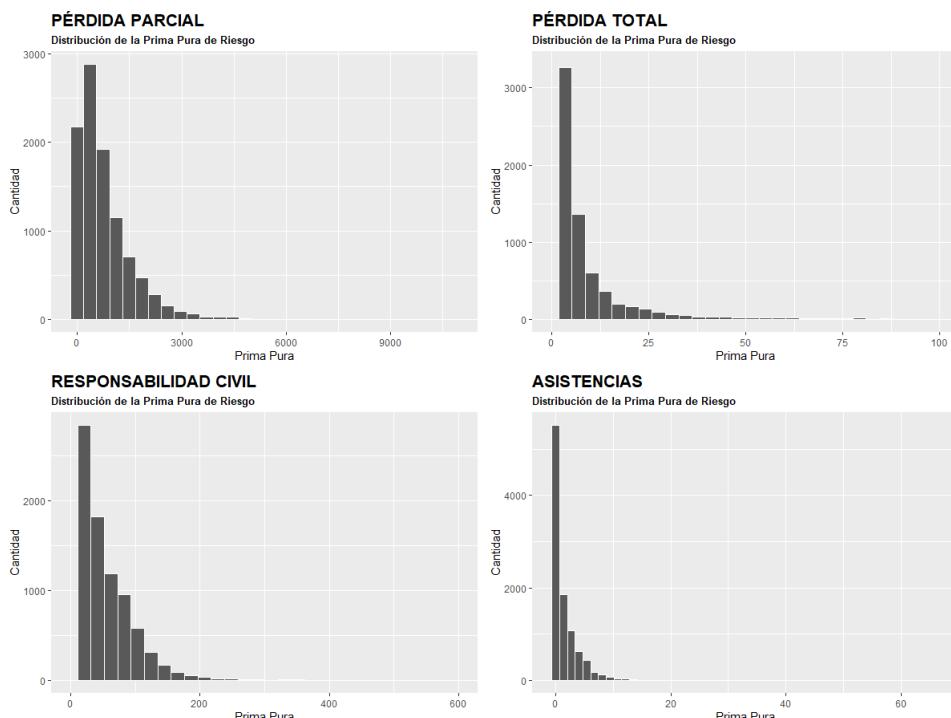
$$f_L(l|\mu, \delta, \lambda, \theta) = \sum_{y=1}^{\infty} \left[D_1 \left(F_X \left(\frac{l}{y} | \mu, \delta \right), F_Y(y|\lambda) | \theta \right) - D_1 \left(F_X \left(\frac{l}{y} | \mu, \delta \right), F_Y(y-1|\lambda) | \theta \right) \right] \cdot \frac{1}{y} f_X \left(\frac{l}{y} | \mu, \delta \right)$$

como se puede observar, la función de pérdida te pide las distribuciones marginales F_X y F_Y , además de los parámetros $\mu, \delta, \lambda, \theta$, para luego aplicar la siguiente integral que nos dará el valor esperado de la pérdida o la Prima Pura de Riesgo:

$$PPR = E(L) = \int_0^{\infty} l \times f_L(l|\mu, \delta, \lambda, \theta) dl$$

Debido a la alta capacidad computacional requerida, tomamos una muestra de 10,000 vehículos asegurados de nuestra base de datos para visualizar la distribución de la Prima Pura de Riesgo por cobertura, por lo que obtenemos:

FIGURA N° 4.7: Distribución de la Prima Pura de Riesgo



Dado que hemos tomado una muestra, necesitaremos ajustar cada distribución a una distribución de probabilidad conocida como la Gamma, para ello estimaremos los parámetros para cada una mediante Máxima Verosimilitud. Hasta este punto ya tenemos las primas pura de riesgo por cobertura, sin embargo, como hemos indicado en nuestras hipótesis, en la ocurrencia del siniestro suelen activarse más de una cobertura por lo que suele existir cierto grado de dependencia entre las pérdidas por cobertura, que son cubiertas por el pago de una sola prima. En ese sentido nuestro enfoque es el siguiente:

FIGURA N° 4.8: Ajuste a una distribución Gamma - Estimación de parámetros

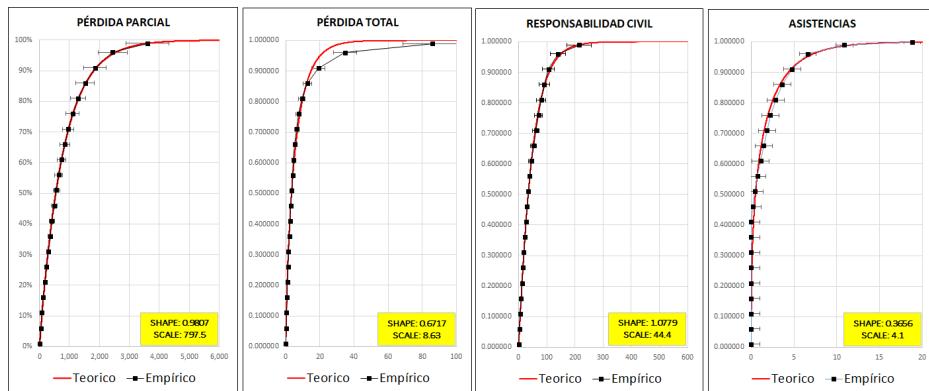
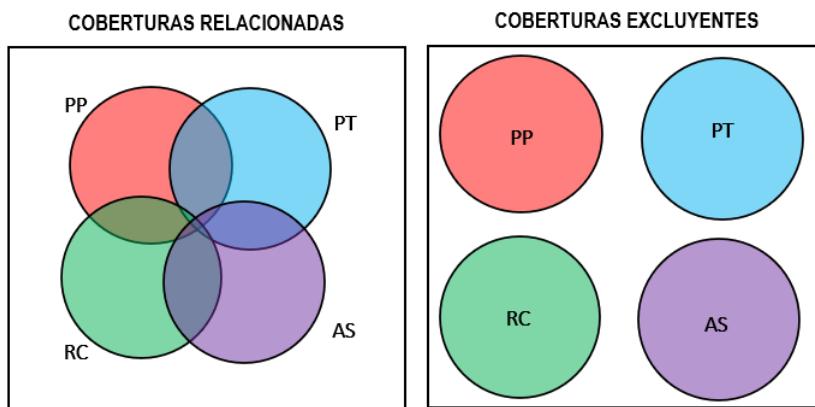


FIGURA N° 4.9: Formas de agregación de las primas por cobertura



El cuadro de la derecha de la [Figura N° 4.9](#) indica que las coberturas son excluyentes entre sí, por lo que la prima final agregada será la suma simple de cada una. Este supuesto práctico es generalmente asumido en el ejercicio real de las aseguradoras.

Parte de nuestra propuesta metodológica es asumir cierto grado de relación entre las primas de cada cobertura, como el cuadro de la izquierda de la [Figura N° 4.9](#), siendo no necesariamente la estructura relacional la que se muestra, es solo referencial. En ese sentido, la agregación de las primas de riesgo por cobertura no se limita a una suma simple, sino a una agregación de riesgos bivariados y dependientes.

El esquema que utilizaremos para agregar los riesgos por cobertura será el siguiente:

Como tenemos cuatro coberturas por agregar, existen seis formas de parejas bivariadas que se pueden formar, como debemos tomar solo una de ellas, nuestro criterio será elegir aquella pareja cuya grado de concordancia sea el mayor mediante el τ de Kendall. Los grados de asociación fueron los siguientes:

De acuerdo a los grados de asociación obtenidos iniciaremos la agrupación bivariada, bajo el criterio de mayor asociación, la primera pareja serán las coberturas de Pérdida Parcial y

FIGURA N° 4.10: Esquema de agregación de riesgos por cobertura

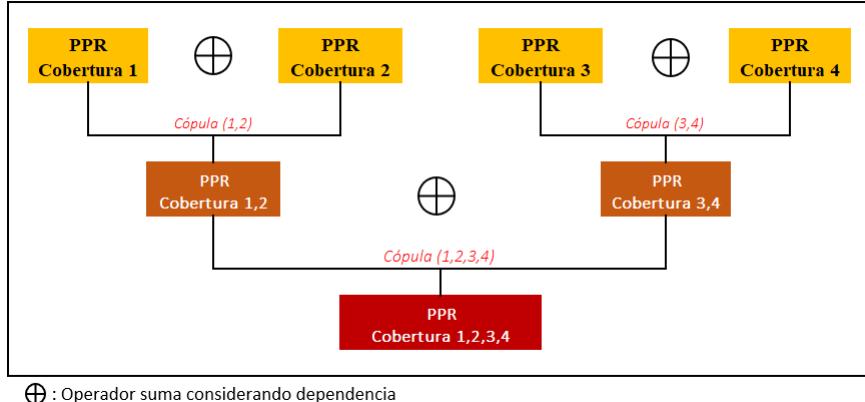


TABLA N° 4.24: Grados de dependencia bivariado - Tau de Kendall

Nº	Pareja Bivariada	τ	Pareja Bivariada	τ
1	PP y PT	56.75%	RC y AS	17.76%
2	PP y RC	26.29%	PT y AS	17.71%
3	PP y AS	14.67%	PT y RC	32.03%

Pérdida Total, y la segunda pareja serán las coberturas de Responsabilidad Civil y Asistencia.

Entonces para agregar la prima pura de riesgo de la Pérdida Parcial y la Pérdida Total, necesitamos elegir una familia de cópulas que mejor refleje su asociación, bajo el criterio de mayor *loglikelihood* obtuvimos:

- PP y PT: elegimos la cópula de Clayton con parámetro $\theta = 0.714$ y *loglik* = 1582.
- RC y AS: elegimos la cópula de Clayton con parámetro $\theta = 0.504$ y *loglik* = 424

el detalle de cálculo y códigos en R utilizados se encuentran en el Anexo C.

El siguiente proceso es estimar la prima pura de riesgo bivariada de cada pareja de coberturas, este proceso se realizará mediante simulación montecarlo utilizando la propiedad descrita en la subsección 2.3.1.1., es decir:

- Simulamos q y u aleatoriamente entre 0 y 1.
- Calculamos v según lo indicado en la Tabla 2.9 de la subsección 2.3.1.1.

$$v = \left(1 + u^{-\theta} (q^{\frac{-\theta}{\theta+1}} - 1)\right)^{\frac{-1}{\theta}}$$

- Entonces dado que tenemos los vectores u y v procedemos a calcular la prima pura de riesgo agregada para la Pérdida Parcial y Pérdida Total de la siguiente forma:

$$PPR_{PP+PT} = \text{Gamma}^{-1}(u | shape = 0.098, scale = 797.5) + \text{Gamma}^{-1}(v | shape = 0.672, scale = 8.63)$$

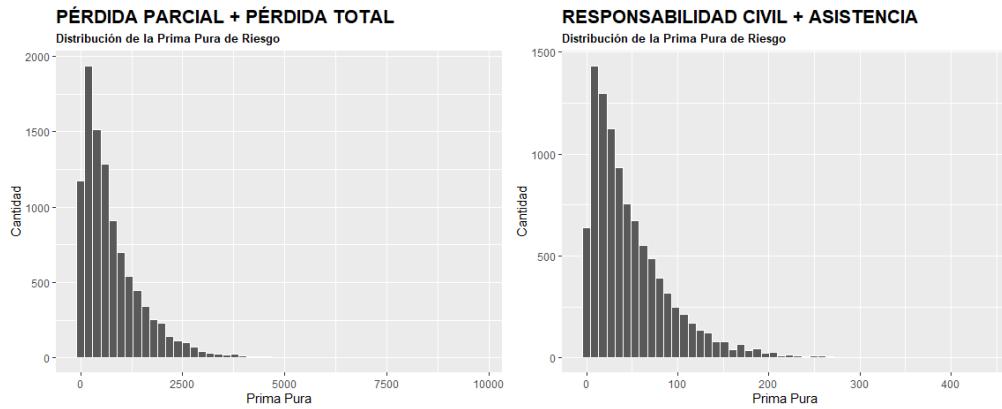
- Para las coberturas Responsabilidad Civil y Asistencias realizamos el mismo procedimiento, y como la cópula clayton fue la de mayor verosimilitud la forma de hallar v es la misma, entonces la prima pura de riesgo agregada es:

$$PPR_{RC+AS} = \text{Gamma}^{-1}(u | \text{shape} = 1.077, \text{scale} = 44.45) + \text{Gamma}^{-1}(v | \text{shape} = 0.366, \text{scale} = 4.13)$$

Observar que los parámetros de *shape* y *scale* son los estimados e indicados en la [Figura N° 4.8](#).

Entonces la distribución de la prima pura de riesgo de cada par de coberturas se muestran en la siguiente figura:

FIGURA N° 4.11: Distribución de la Prima de Riesgo Bivariada



Finalmente, procederemos a agregar las primas bivariadas de cada distribución considerando la relación de dependencia que pueden tener, para ello calcularemos la familia cópula que mejor se ajuste de acuerdo al criterio del *loglink*, en este caso la familia cópula elegida fue la Gumbel con parámetro $\theta = 1.005$, $loglink = 1.0163$ y un grado de dependencia de $\tau = 0.53\%$. El cálculo se detalla en código R en el Anexo C.

Agregamos las primas de riesgo bajo dependencia siguiendo los siguientes pasos:

- Simulamos q y u aleatoriamente entre 0 y 1.
 - Calculamos v según lo indicado en la Tabla 2.9 de la Subsección 2.3.1.1. para la familia Gumbel
- $$v = \exp\left(-\left((-lnqu)^{\theta} - (-lnu)^{\theta}\right)^{\frac{1}{\theta}}\right)$$
- Entonces dado que tenemos los vectores u y v procedemos a calcular la prima pura de riesgo agregada de la siguiente manera:

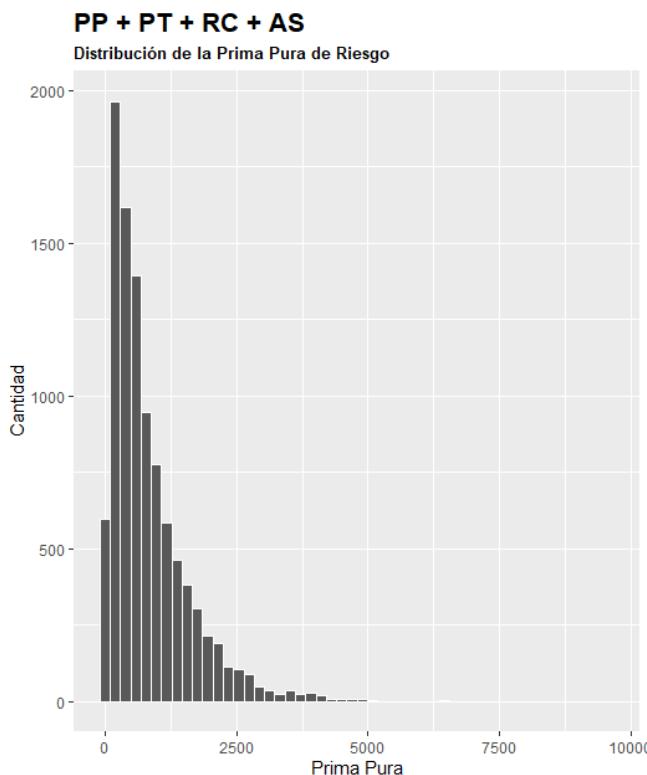
$$PPR_{PP+PT+RC+AS} = F_{PP+PT}^{-1}(u) + F_{RC+AS}^{-1}(v)$$

donde $F(PP + PT)$ y $F(RC + AS)$ son las distribuciones de probabilidad acumulativa

ladas de la Figura N° 4.10.

La distribución de probabilidad de la prima pura de riesgo conjunta de las 4 coberturas consideradas en este trabajo es la siguiente:

FIGURA N° 4.12: Distribución de la Prima de Riesgo Multivariada Pérdida Parcial + Pérdida Total + Responsabilidad Civil + Asistencias



4.3 Interpretación y discusión de los resultados

El presente trabajo ha desarrollado una mixtura entre aspectos que se desarrollan en la práctica diaria del *pricing* de los departamentos actuariales, también se utilizaron tópicos de estudios realizados por los diversos autores mencionados en este trabajo y la propuesta metodológica de llegar a la prima de riesgo multicobertura considerando la dependencia entre riesgos.

Como hemos visto en el desarrollo de este trabajo, los modelos de regresión generalizados marginales de la frecuencia y de la severidad son la base del desarrollo del modelo conjunto, es por ello qué en términos de **Robustez**, estos modelos tienen coeficientes altamente significativos, como se muestran en el Anexo C, además, de acuerdo a los resultados del índice de GINI se demostró que los riesgos que se están analizando no todos son buenos o malos, sino que es posible segmentarlos. Por lo tanto, en este punto los resultados son acorde a la práctica actuarial en nuestra localidad.

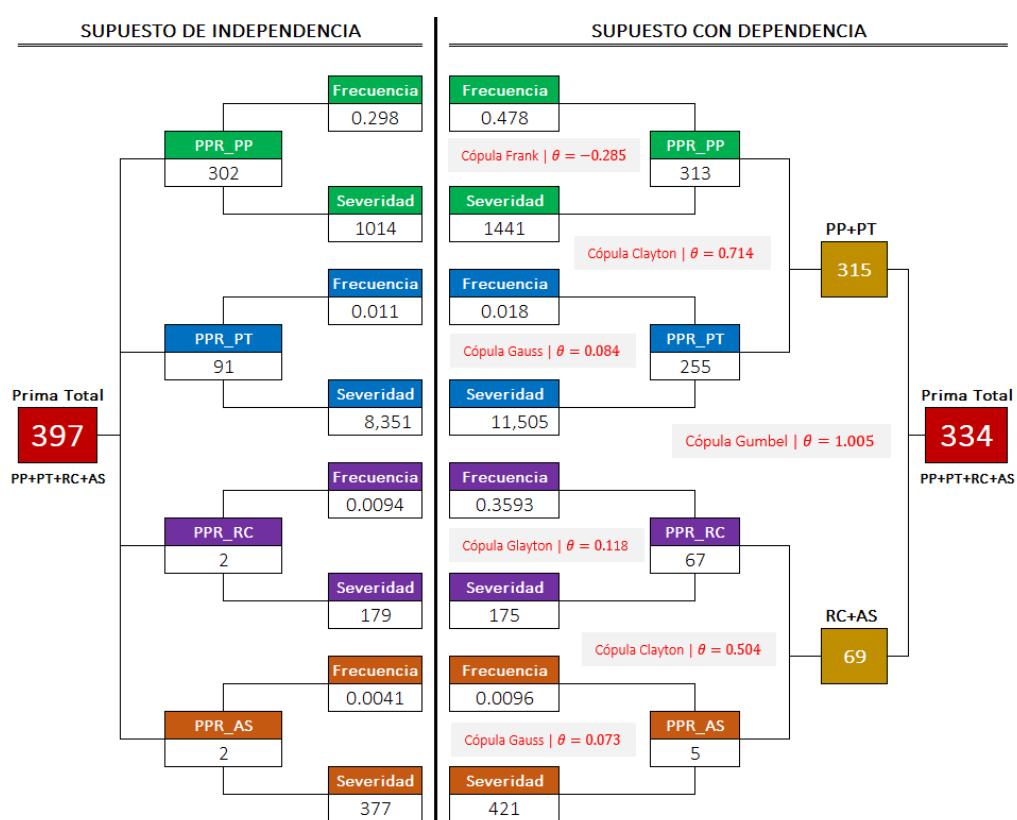
De los aspectos considerados, el más importante ha sido el de utilizar familias de cópulas, en particular las propiedades de la cópula-mixta, para la construcción de modelos conjuntos de probabilidad y poder medir la dependencia entre los riesgos. Los resultados obtenidos en la [Tabla N° 4.20](#) son concordantes con los resultados obtenidos por [Kholifah et al. \(2019\)](#) y por [Krämer et al. \(2013\)](#) en sus respectivos trabajos. Un punto interesante fueron los resultados de los indicios de dependencia obtenidos por el Chi-Plot y el K-plot, ya que diversos trabajos muestran resultados bastante suavizados, sin embargo, del hecho de que los datos provienen de un comportamiento real, los gráficos Chi-Plot y K-Plot mostraron indicios de dependencia con ciertos patrones complejos de dependencia, para entender mejor esto sugiero revisar directamente el trabajo realizado por [Fisher y Switzer \(2001\)](#) ya que se explican diversas casuísticas en los resultados, y no de otros autores que toman su trabajo y se limitan a los casos simples. En este sentido, los resultados de **dependencia** han sido concordantes con la literatura revisada.

Referente a la propuesta metodologica de agregación de las primas de riesgo de cada cobertura, considerando la dependencia entre las mismas, parte del supuesto que la agregación usual y práctica que se realiza en la industria, ésta sobreestima la prima pura de riesgo, según el enfoque mostrado en la [Figura N° 4.9](#). Para comprobar ello he realizado un cotizador que calcule la prima pura de riesgo de acuerdo a las características que pudiera tener un asegurado y su vehículo, y nos permite comparar las primas puras de riesgo bajo el supuesto de independencia y dependencia, considerado todos los parámetros estimados mostrados en la sección de análisis.

Una comparación se muestra en la [Figura N° 4.13](#) de acuerdo a las siguientes variables:

Tipo Póliza	Canal	Edad	Sexo	Tipo vehículo	Marca Auto	Antiguedad	Valor Vehículo	Uso vehículo	Zona
Individual	Bróker	25	Femenino	Liviano	Toyota	0	20,000	Particular	A3

FIGURA N° 4.13: Prima Pura de Riesgo: Independencia vs. Dependencia



CONCLUSIONES

Se ha desarrollado una propuesta metodológica alternativa que incluyeron aspectos teóricos mostrados en las bases teóricas e implementados en la práctica mediante un cotizador. La aplicación sobre una base de datos con características reales demostraron que las hipótesis planteadas se cumplen. Finalmente luego de analizar los resultados obtenidos en el presente trabajo, se presentan las siguientes conclusiones:

1. Mediante las pruebas gráficas se pudo demostrar que existe un grado de asociación en la Frecuencia y la Severidad. Se observó que patrón dependencia no es lineal, al contrario, tiene patrones complejos de dependencia a lo largo del cuerpo de la distribución condicional.
2. La construcción de un modelo conjunto de la Frecuencia y de la Severidad fue posible utilizando propiedades de cópula mixta. Los coeficientes de regresión conjunta estimados son razonables, no son muy distintos al modelo marginal estimado, se entiende que las diferencias halladas se explican por efecto de la dependencia considerado.
3. Se pudo medir el grado de dependencia bajo el concepto de concordancia entre variables aleatorias, principalmente utilizando la Tau de Kendall.
4. La propuesta metodológica de agregación bivariada de las primas puras de riesgo por cobertura, para llegar a la prima total, nos pudo demostrar la hipótesis de sobre estimación. Por lo tanto, esta propuesta calcula una prima más realista acorde a las características de riesgo del asegurado.

RECOMENDACIONES

La metodología desarrollada en cierto sentido estima una prima pura de riesgo más compacta, sin embargo, en el transcurso se hallaron ciertas limitaciones, por lo que el objetivo de las siguientes recomendaciones es mejorar o complementar el trabajo realizado. Por lo tanto, las recomendaciones son las siguientes:

1. Una vez estimado los parámetros de las regresiones conjuntas, la simulación y/o predicción de valores ajustados al modelo suele requerir un alto performance a nivel computacional. Si bien la información disponible suele estar cercano al millón de datos, hubo en algunos puntos del trabajo donde se utilizó un muestra representativa para realizar algunas estimaciones, como por ejemplo el cálculo de los errores estándar de los coeficientes del modelo al momento de estimar su matriz Hessiana.
2. Un interesante complemento al trabajo es hallar la función de densidad conjunta para 4 riesgos (referente a la cantidad de coberturas) utilizando las propiedades de las Vine Cúpulas, si bien fueron mencionadas en las bases teóricas, el presente trabajó se basó en otra alternativa también mostrada.
3. Una extensión al presente trabajo de tesis es calcular el índice de siniestralidad o *Loss Ratio* con el modelo multivariado conjunto estimado, y poder medir el impacto sobre la siniestralidad real, lo que se reflejaría en un aumento en el margen de utilidad. Esto no pudo ser posible, debido a que la información pública incluye datos de diferentes empresas de seguro y el valor de la prima comercial incluía aspectos de impuestos que no se pudieron separar en el proceso para poder estimar la prima devengada por periodo.

REFERENCIAS BIBLIOGRÁFICAS

- Aas, K., Czado, C., Frigessi, A., y Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2), 182–198.
- Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., y Thandi, N. (2007). *A practitioner's guide to generalized linear models—a foundation for theory, interpretation and application*. Towers Watson.
- Anscombe, F. J. (1973). Graphs in statistical analysis. *The american statistician*, 27(1), 17–21.
- Arbenz, P., Hummel, C., y Mainik, G. (2012). Copula based hierarchical risk aggregation through sample reordering. *Insurance: Mathematics and Economics*, 51(1), 122–133.
- Brockman, M. J., y Wright, T. S. (1992). Statistical motor rating: making efficient use of your data. *J. Inst. Actuar.*, 119, 457–543.
- Burden, R., Faires, D., y Burden, A. (2015). *Numerical analysis* (10a edición ed.). Cengage Learning.
- Czado, C., Kastenmeier, R., Brechmann, E. C., y Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 4, 278–305.
- De Jong, P., Heller, G. Z., y cols. (2008). Generalized linear models for insurance data. Cambridge Books.
- de Leon, A. R., y Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 30(2), 175–185.
- Embrechts, P. (1999). Correlation: pitfalls and alternatives. *Risk Magazine*, 69–71.
- Fisher, N., y Switzer, P. (2001). Graphical assessment of dependence: Is a picture worth 100 tests? *The American Statistician*, 55(3), 233–239.
- Frees, E. W., Lee, G., y Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risk*, 4(4), 1–36.
- Garrido, J., Genest, C., y Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70, 205–215.
- Genest, C., y Boies, J.-C. (2003). Detecting dependence with kendall plots. *The American Statistician*, 57(4), 275–284.
- Genest, C., y Neslehova, J. (2007). A primer on copulas for count data. *The Astin Bulletin*, 37, 475–515.
- Joe, H. (2014). *Dependence modeling with copulas*. CRC press.

- Kaas, R., Goovaerts, M., Dhaene, J., y Denuit, M. (2008). *Modern actuarial risk theory using r*. Springer Berlin, Heidelberg.
- Kholifah, A. R., Lestari, D., y Devila, S. (2019). Premium calculation using marginal generalized linear model combined with copula. En (Vol. 2168). doi: 10.1063/1.5132462
- Krämer, N., Brechmann, E. C., Silvestrini, D., y Czado, C. (2013). Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, 53(3), 829–839.
- Lee, W., Park, S., y Ahn, J. Y. (2018). Investigating dependence between frequency and severity via simple generalized linear models. *Journal of the Korean Statistical Society*.
- Lee, W., Park, S. C., y Ahn, J. Y. (2019). Investigating dependence between frequency and severity via simple generalized linear models. *Journal of the Korean Statistical Society*, 48(1), 13–28.
- Lundberg, F. (1903). *I. approximerad framställning af sannolikhetsfunktionen. ii. aterför-säkring af kollektivrisker*. Almqvist Wiksell.
- Matos, J., y cols. (2007). Uncertainty treatment in civil engineering numerical models.
- McCullagh, P., y Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall.
- McNeil, A., Frey, R., y Embrechts, P. (2005). *Quantitative risk management: concepts, techniques and tools*. Princeton University Press.
- Nelsen, R. B. (2006). *An introduction to copulas*. Springer.
- Ohlsson, E., y Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Springer.
- Rao, C. R., Shalabh, Toutenburg, H., y Heumann, C. (2010). *Linear models and generalizations*. Springer Berlin, Heidelberg.
- Renshaw, A. E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin: the Journal of the IAA*, 24(2), 265–285.
- Scarsini, M. (1984). On measures of concordance. *Stochastica*, 8(3), 201–218.
- Schepsmeier, U., y Stöber, J. (2014). Derivatives and fisher information of bivariate copulas. *Statistical papers*, 55(2), 525–542.
- Sikov, A., y Cerdá-Hernández, J. (2021). *Teoría de riesgo para seguros*. EDUNI.
- Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. En *Annales de l'isup* (Vol. 8, pp. 229–231).
- Song, P., Li, M., y Yuan, Y. (2009). Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65(1), 60–68.

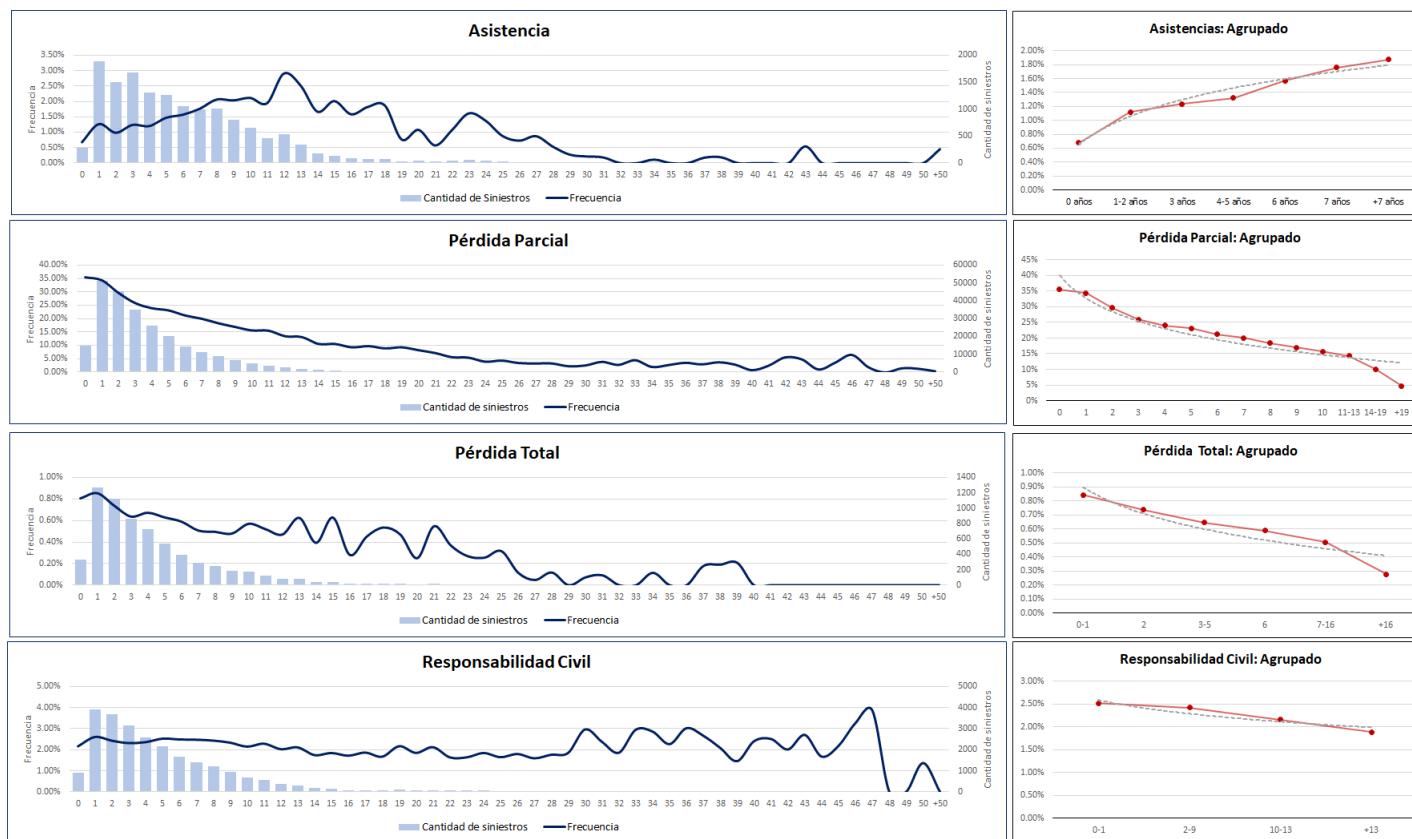
- Song, P. X.-K. (2000). Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2), 305–320.

ANEXOS

ANEXO A: ANÁLISIS UNIVARIADO

Variable: Antiguedad del Vehículo

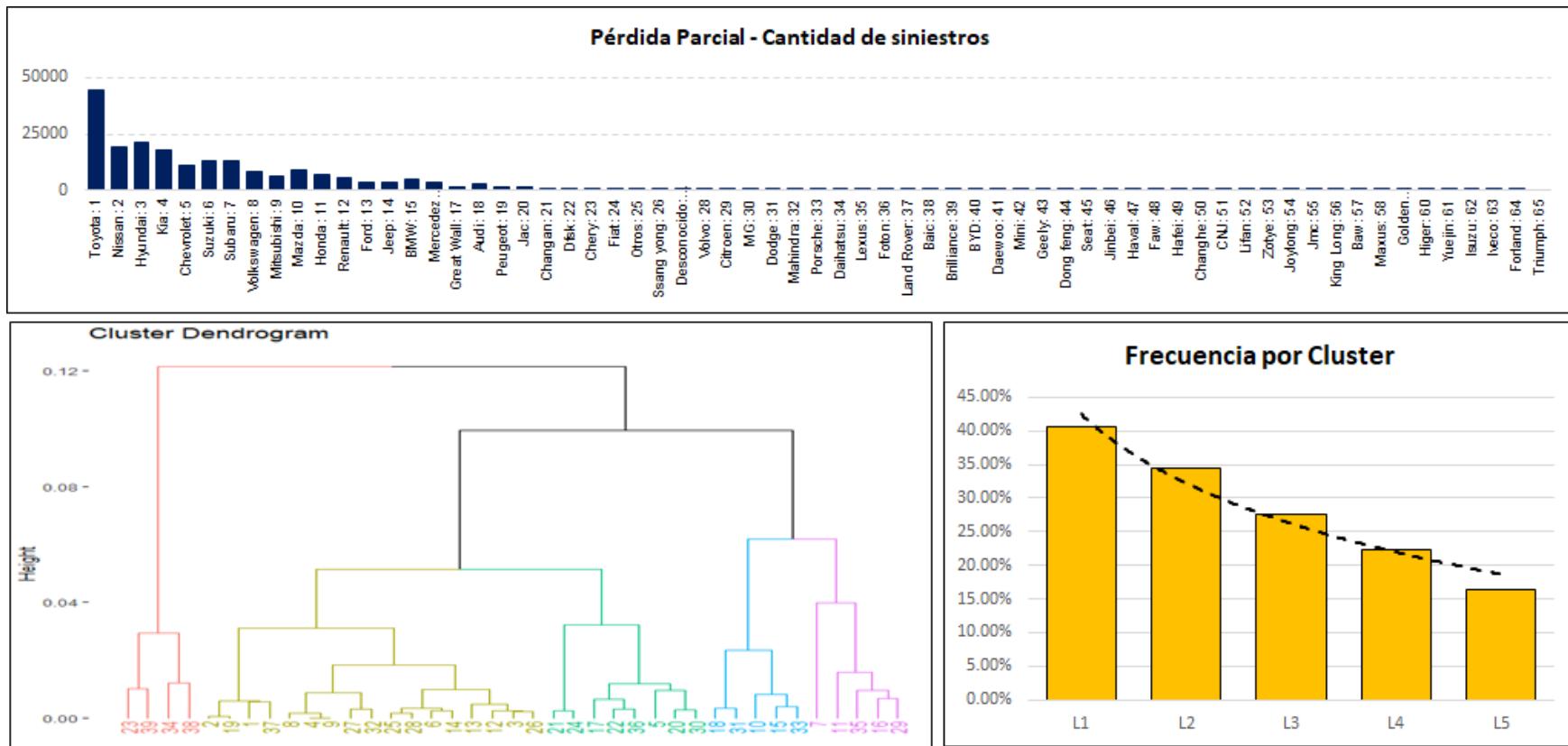
Figure A .1: Agrupación de la variable Antiguedad del vehículo



Variable: Marca de Vehículo

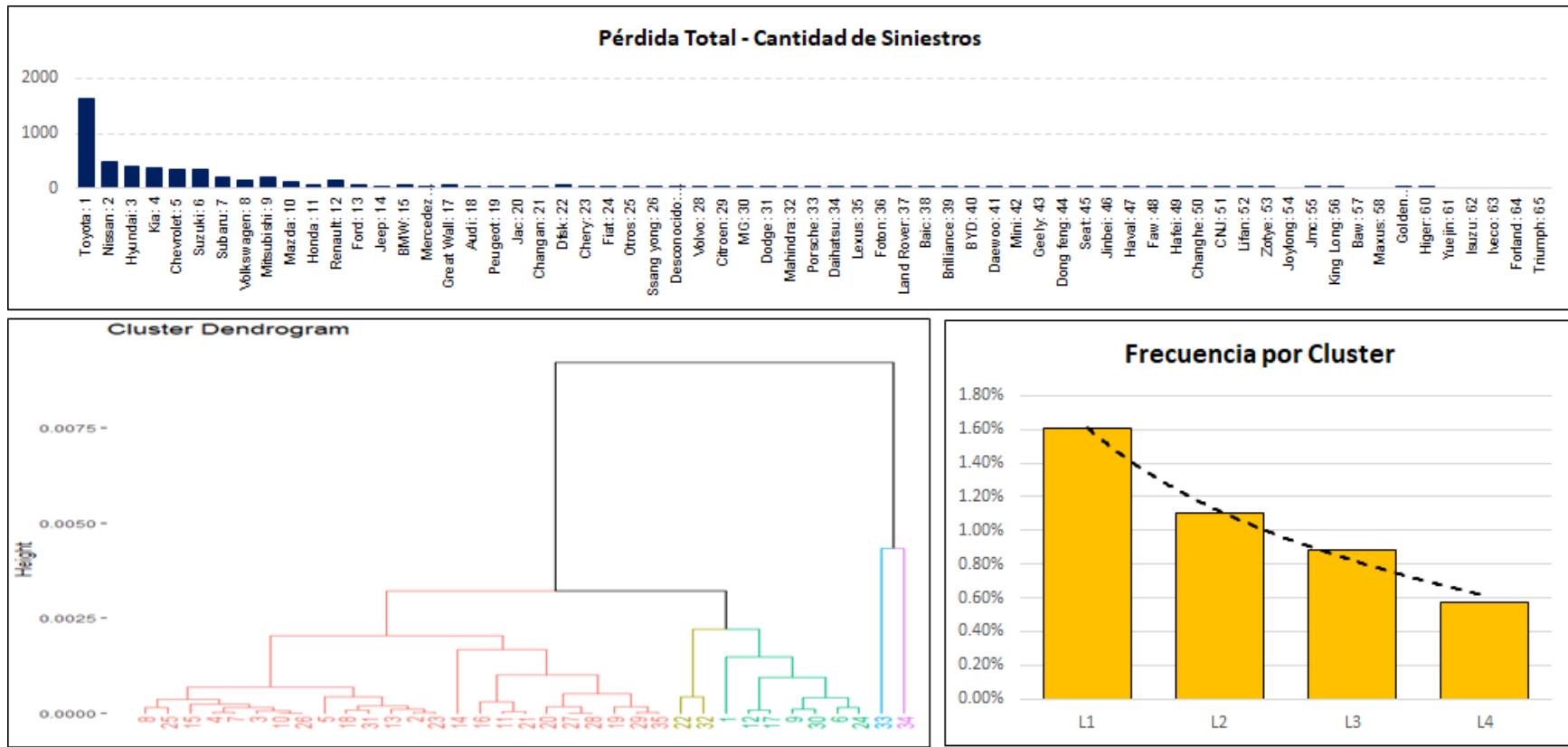
VEHÍCULOS LIVIANOS - PÉRDIDAS PARCIALES

Figure A .2: Agrupación de la Marca Vehículo - Pérdida Parcial - Livianos



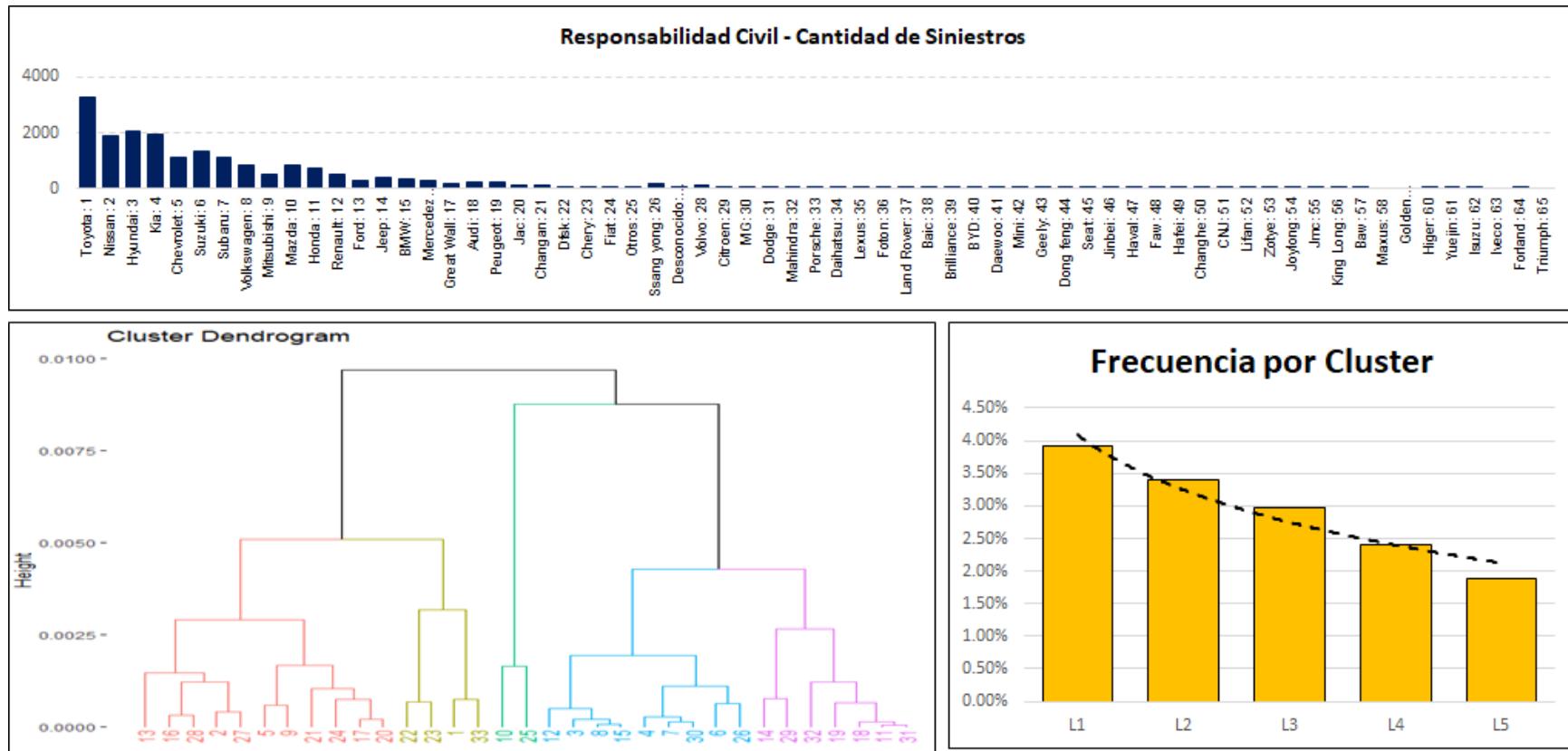
VEHÍCULOS LIVIANOS - PÉRDIDAS TOTALES

Figure A .3: Agrupación de la Marca Vehículo - Pérdida Total - Livianos



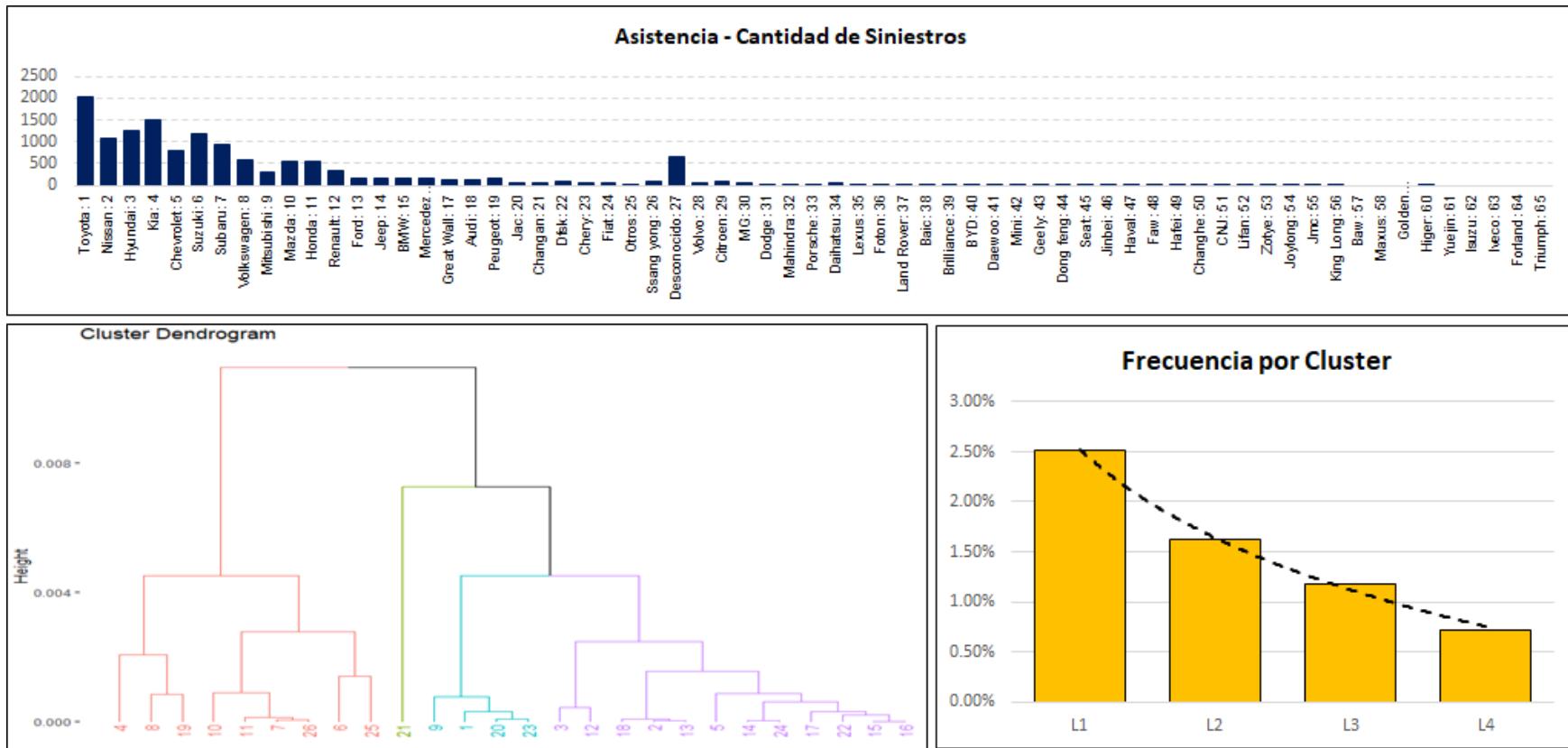
VEHÍCULOS LIVIANOS - RESPONSABILIDAD CIVIL

Figure A .4: Agrupación de la Marca Vehículo - Responsabilidad Civil - Livianos



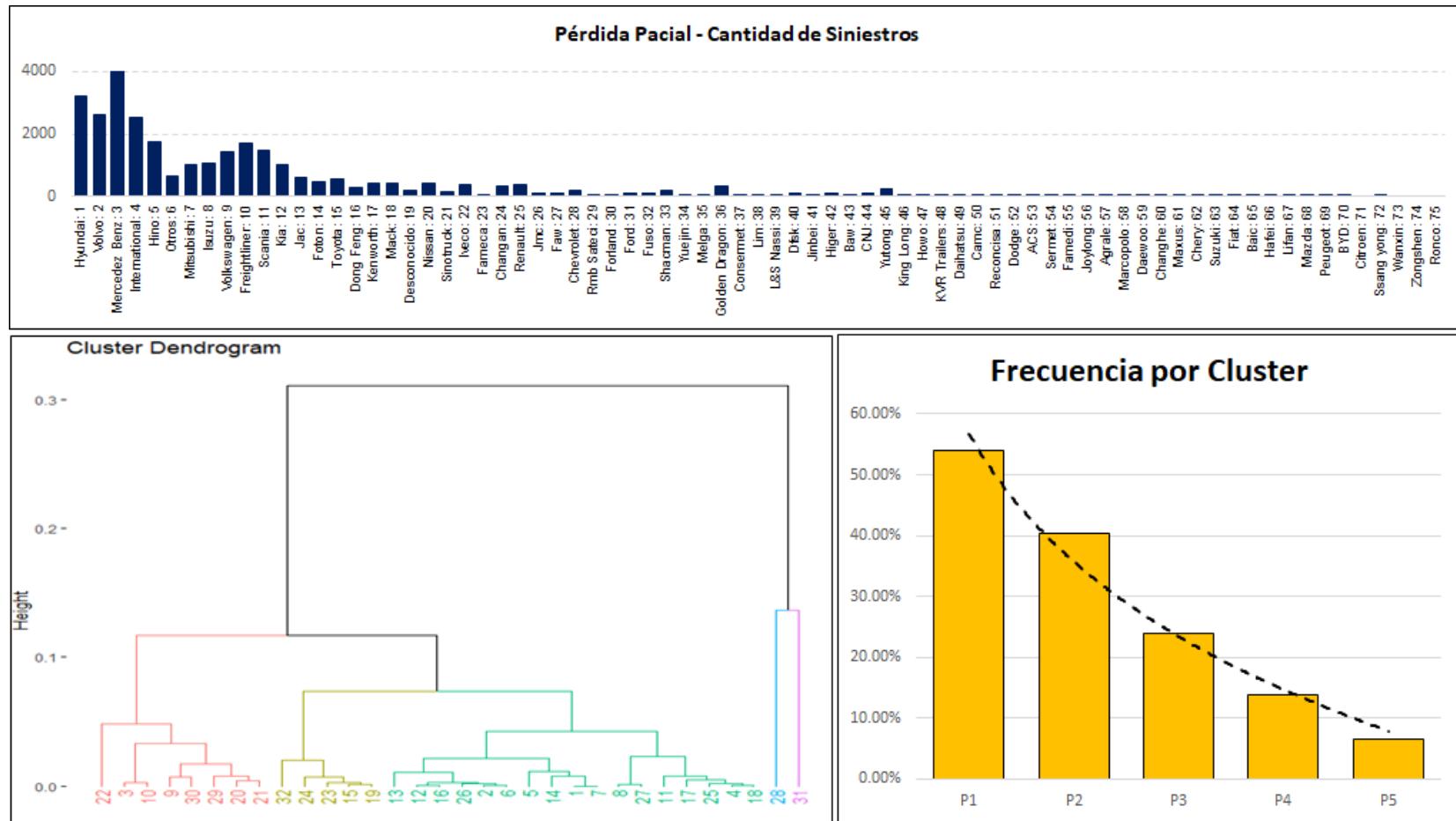
VEHÍCULOS LIVIANOS - ASISTENCIAS

Figure A .5: Agrupación de la Marca Vehículo - Asistencias - Livianos



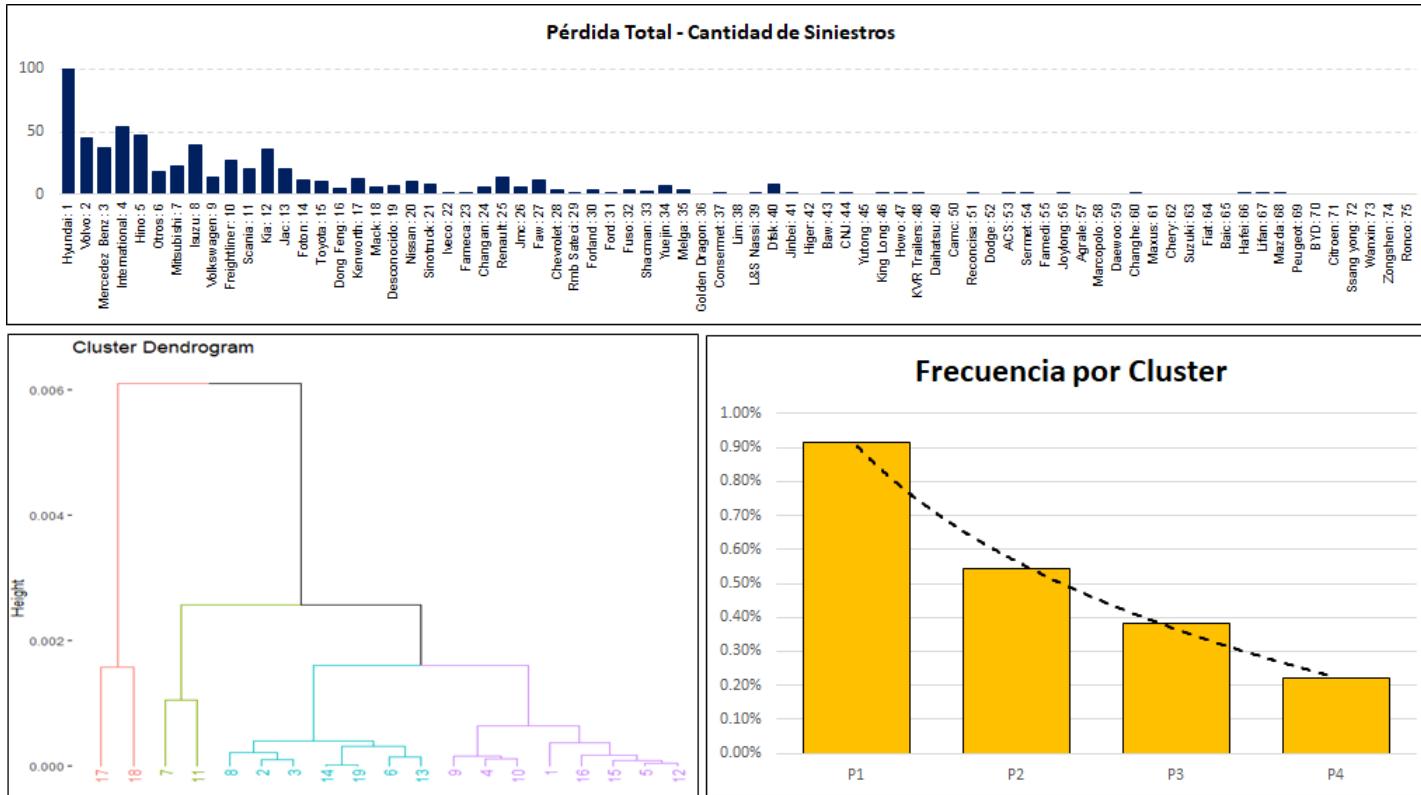
VEHÍCULOS PESADOS - PÉRDIDAS PARCIALES

Figure A .6: Agrupación de la Marca Vehículo - Pérdida Parcial - Pesados



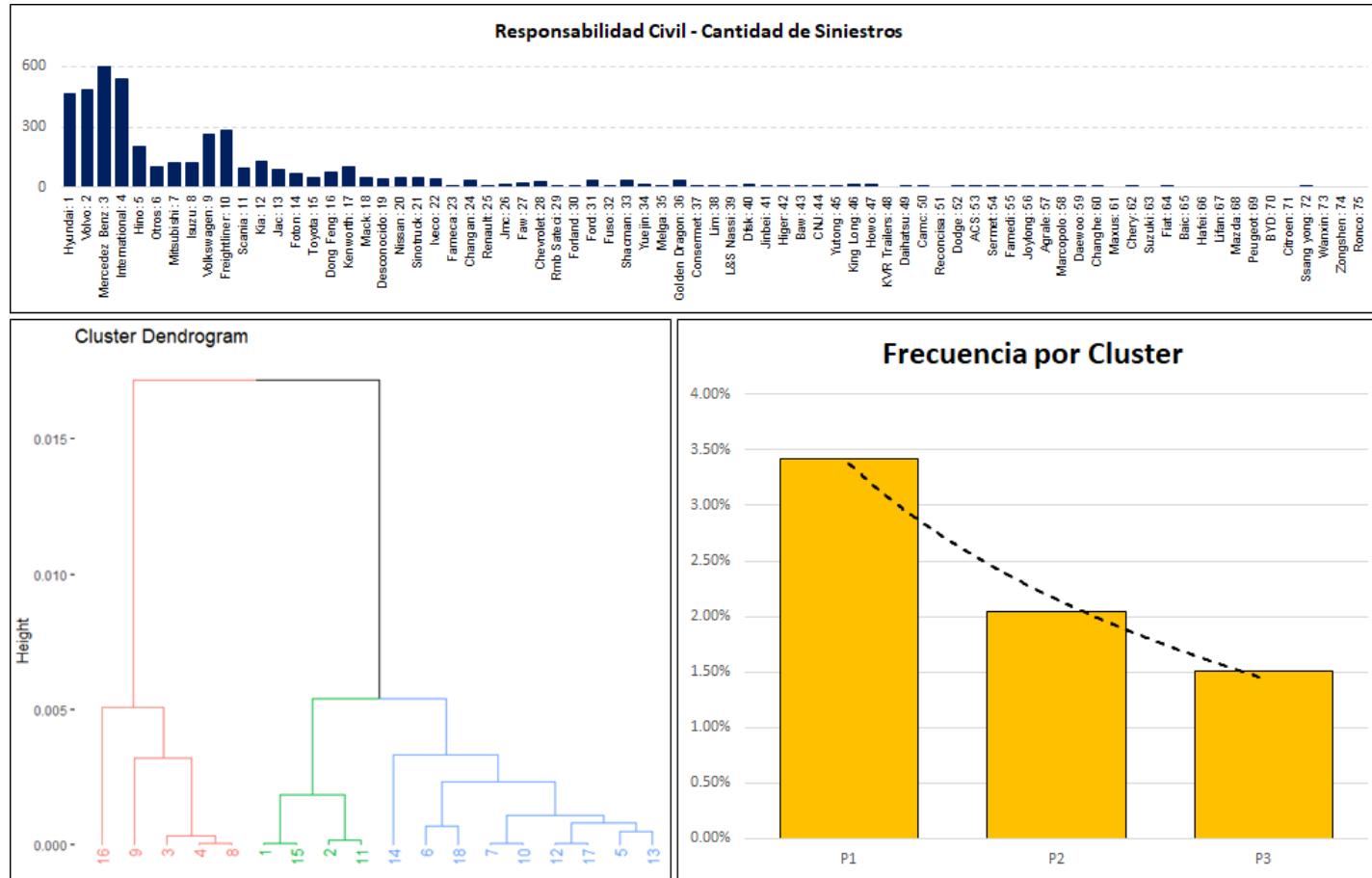
VEHÍCULOS PESADOS - PÉRDIDAS TOTALES

Figure A .7: Agrupación de la Marca Vehículo - Pérdida Total - Pesados



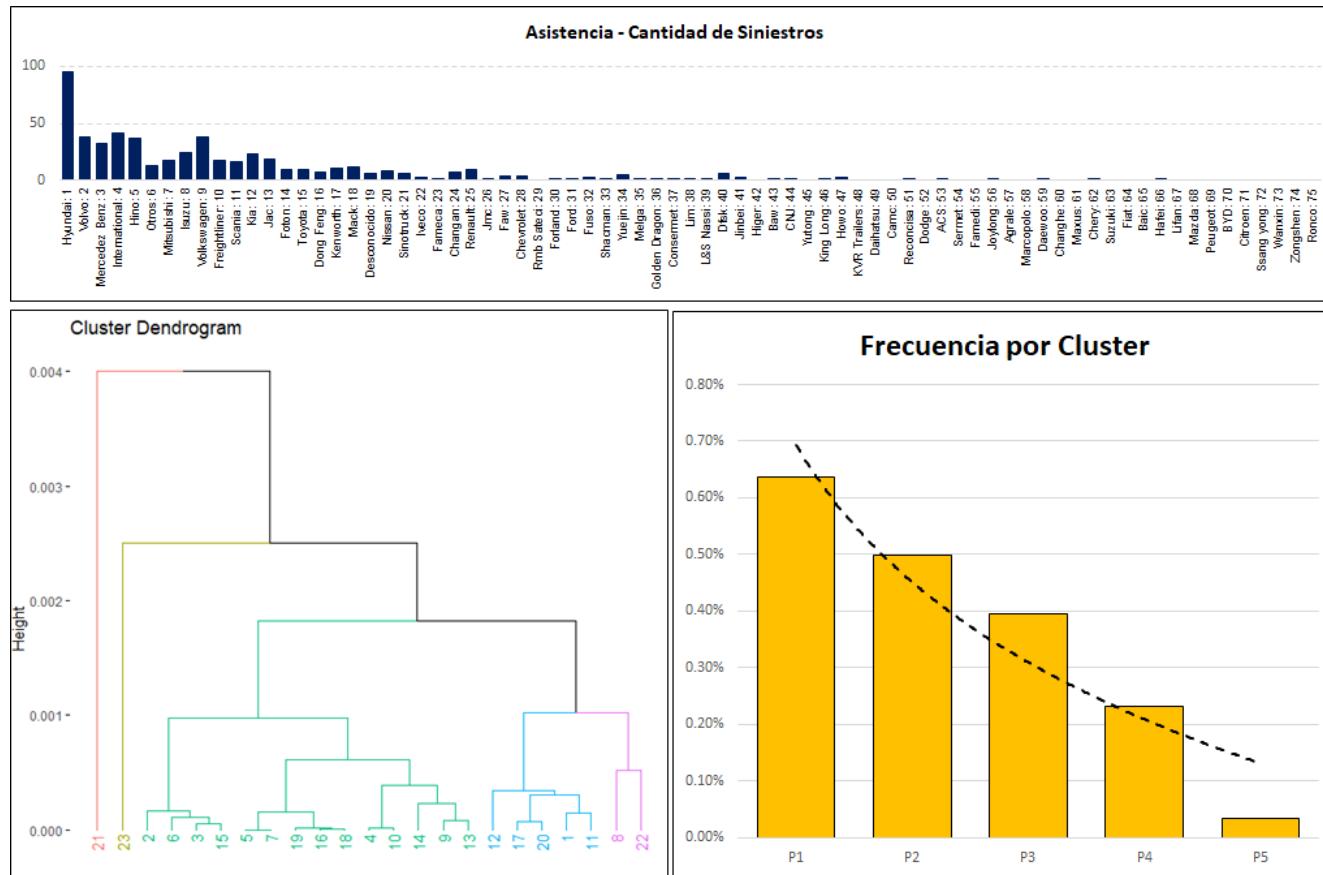
VEHÍCULOS PESADOS - RESPONSABILIDAD CIVIL

Figure A .8: Agrupación de la Marca Vehículo - Responsabilidad Civil - Pesados



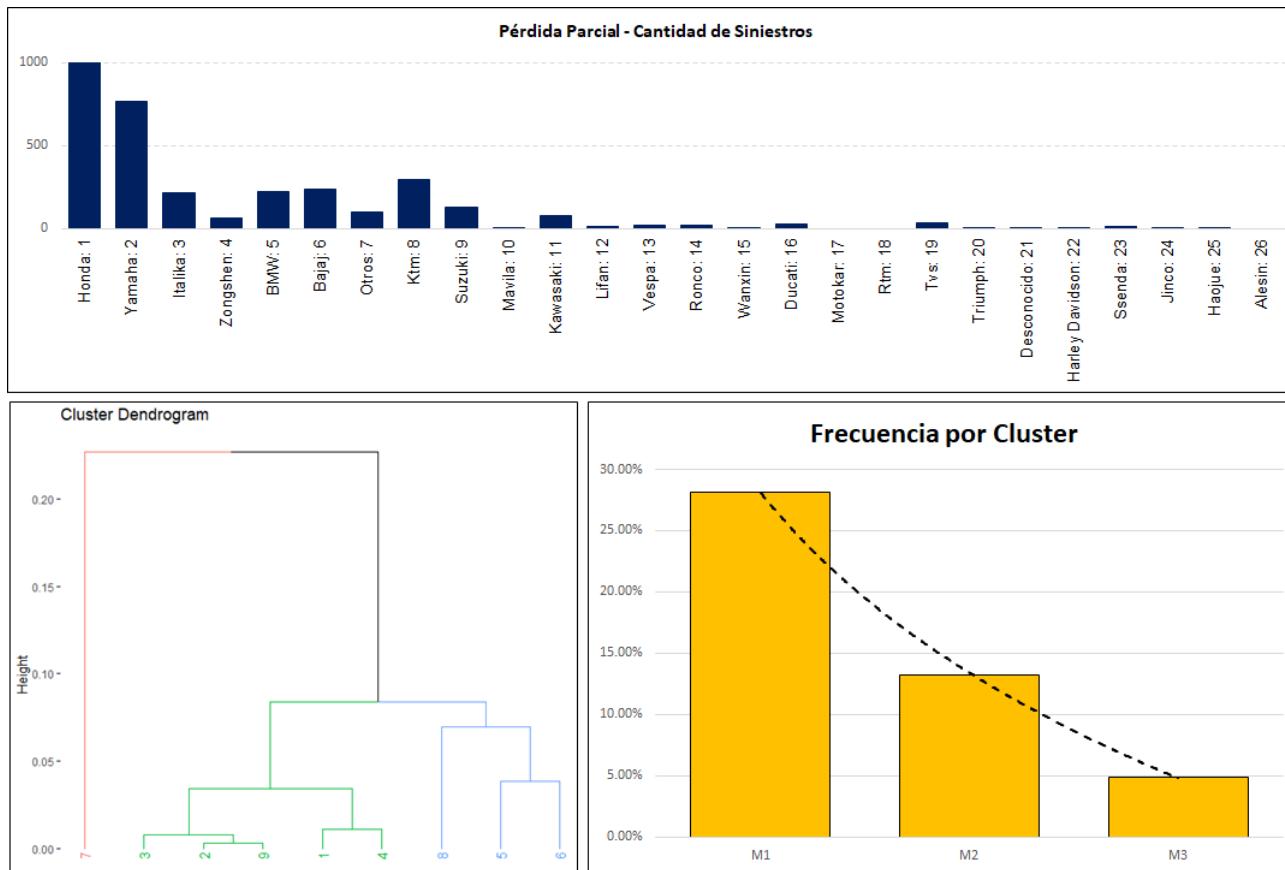
VEHÍCULOS PESADOS - ASISTENCIAS

Figure A .9: Agrupación de la Marca Vehículo - Asistencias - Pesados



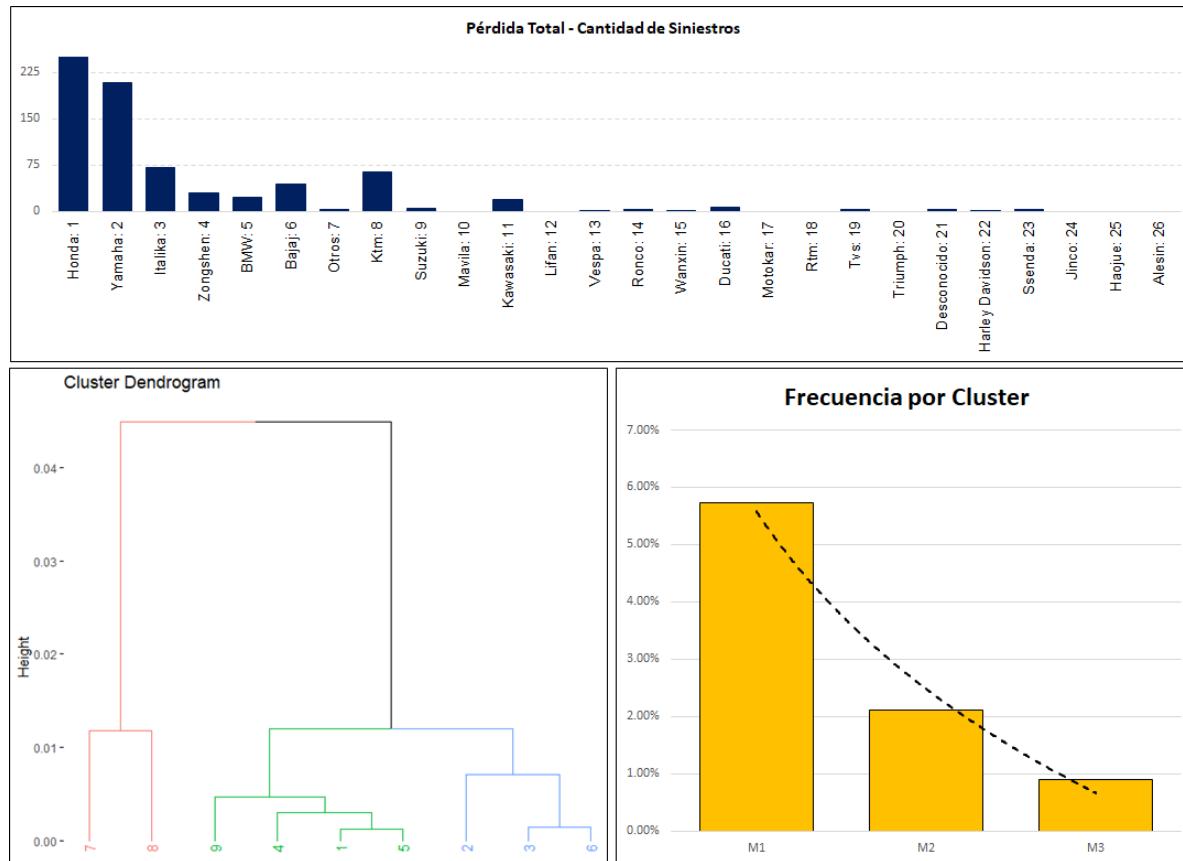
VEHÍCULOS MENORES - PÉRDIDAS PARCIALES

Figure A .10: Agrupación de la Marca Vehículo - Pérdida Parcial - Vehículos Menores



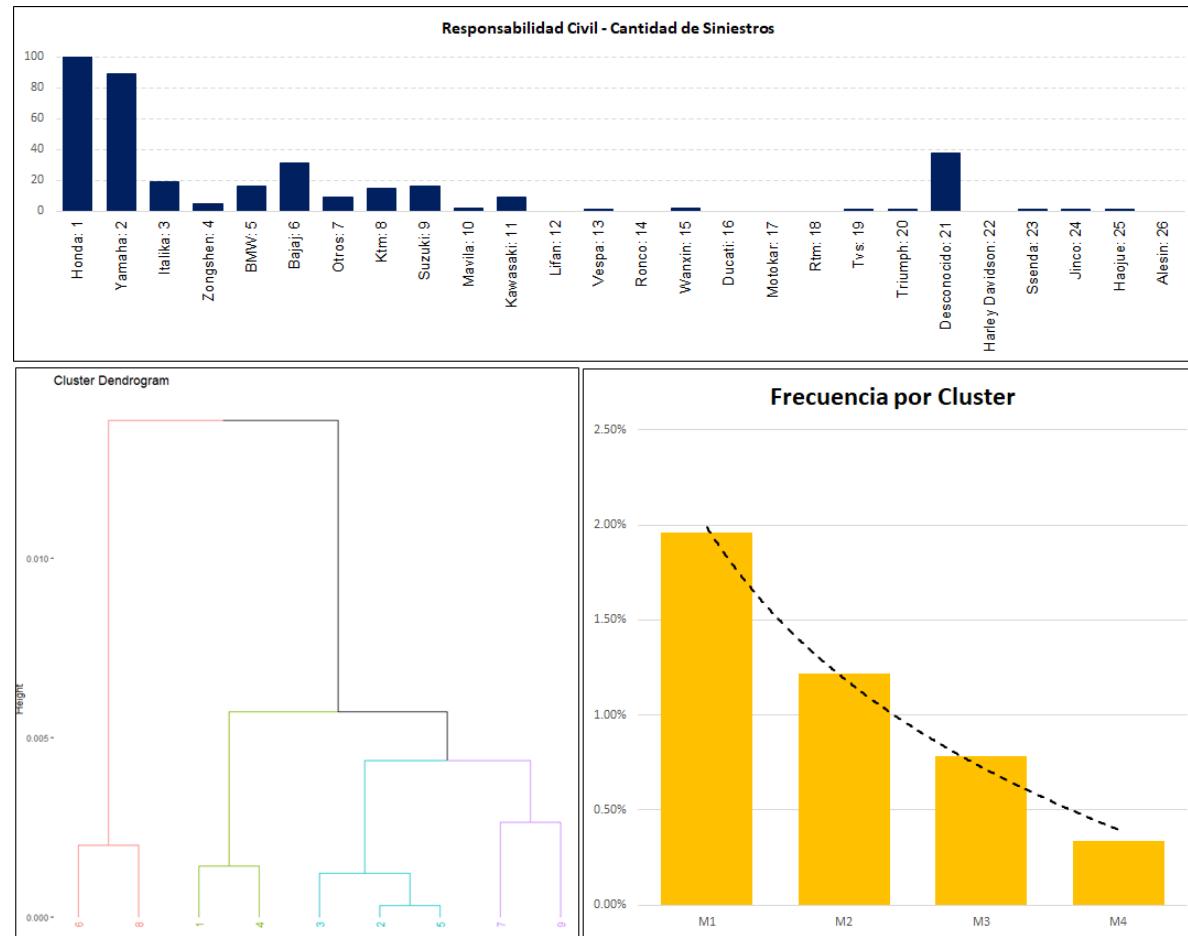
VEHÍCULOS MENORES - PÉRDIDAS TOTALES

Figure A .11: Agrupación de la Marca Vehículo - Pérdida Total - Vehículos Menores



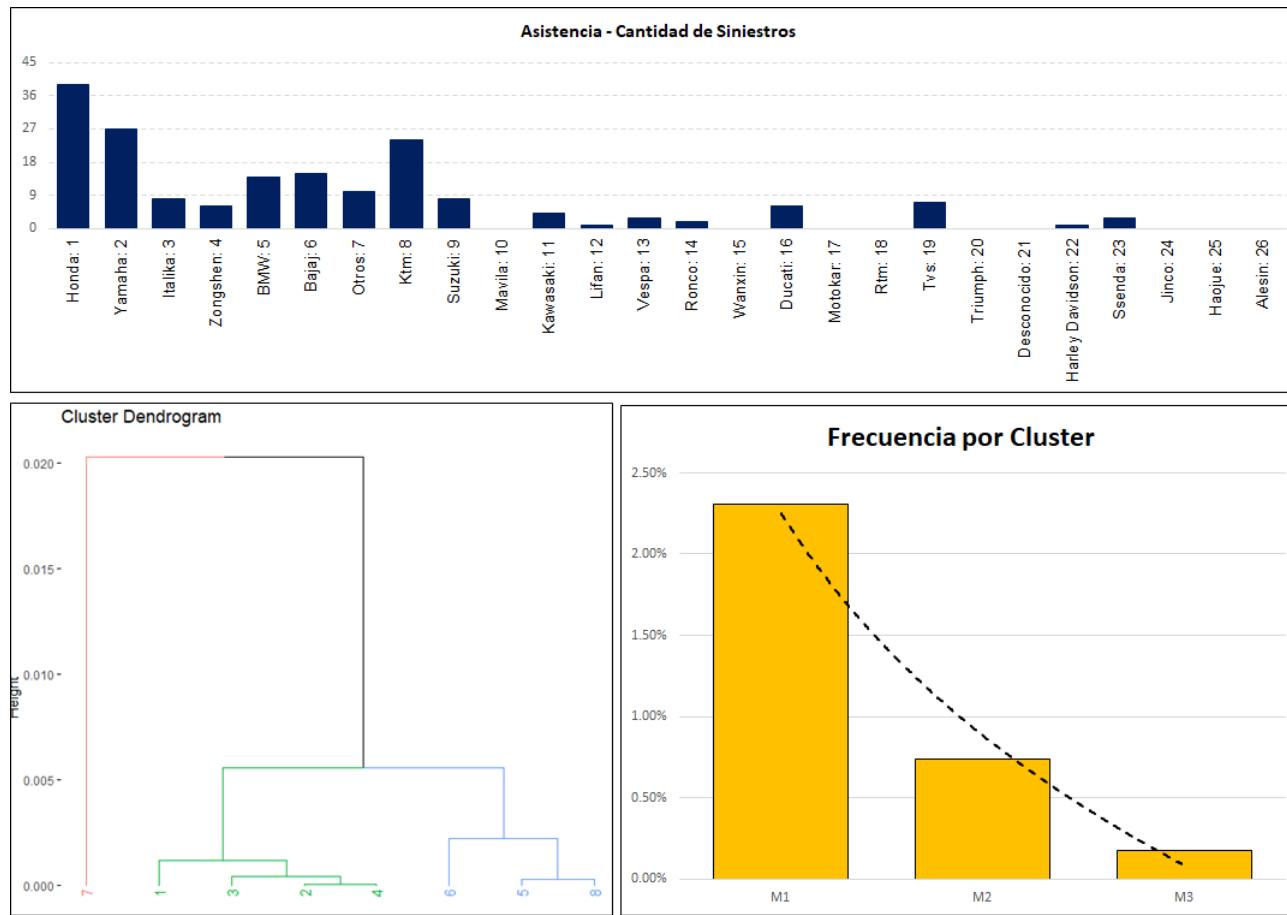
VEHÍCULOS MENORES - RESPONSABILIDAD CIVIL

Figure A .12: Agrupación de la Marca Vehículo - Responsabilidad Civil - Vehículos Menores



VEHÍCULOS MENORES - ASISTENCIAS

Figure A .13: Agrupación de la Marca Vehículo - Asistencias - Vehículos Menores



Variable: Edad del Asegurado

Figure A .14: Edad del asegurado - Sin Agrupar

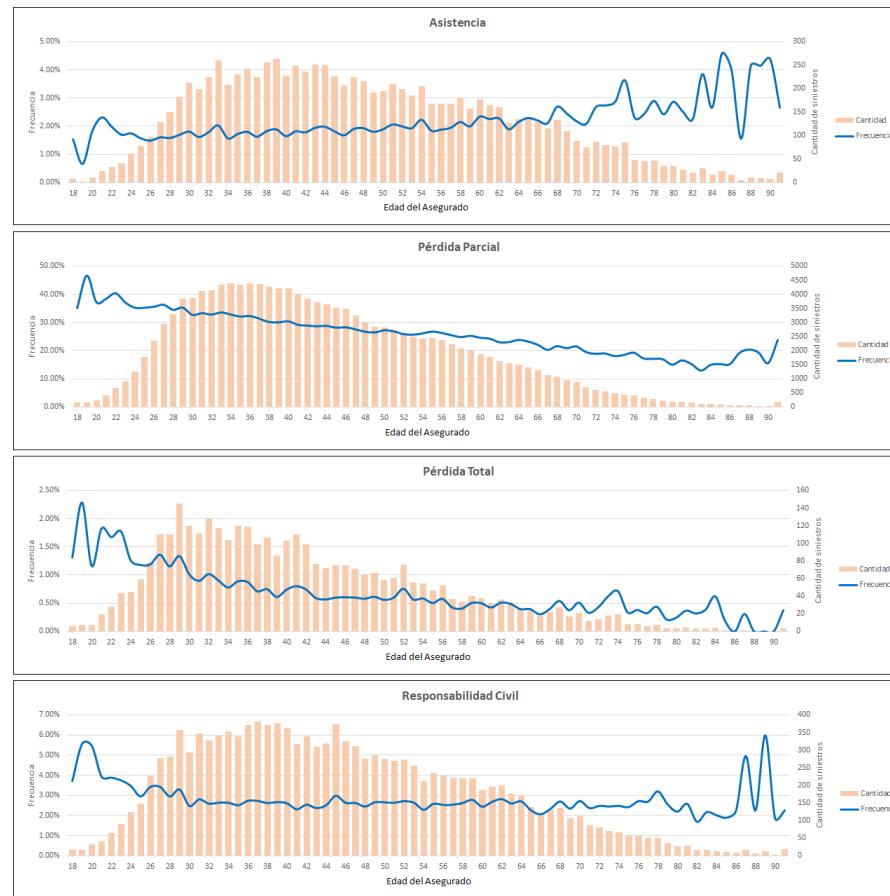
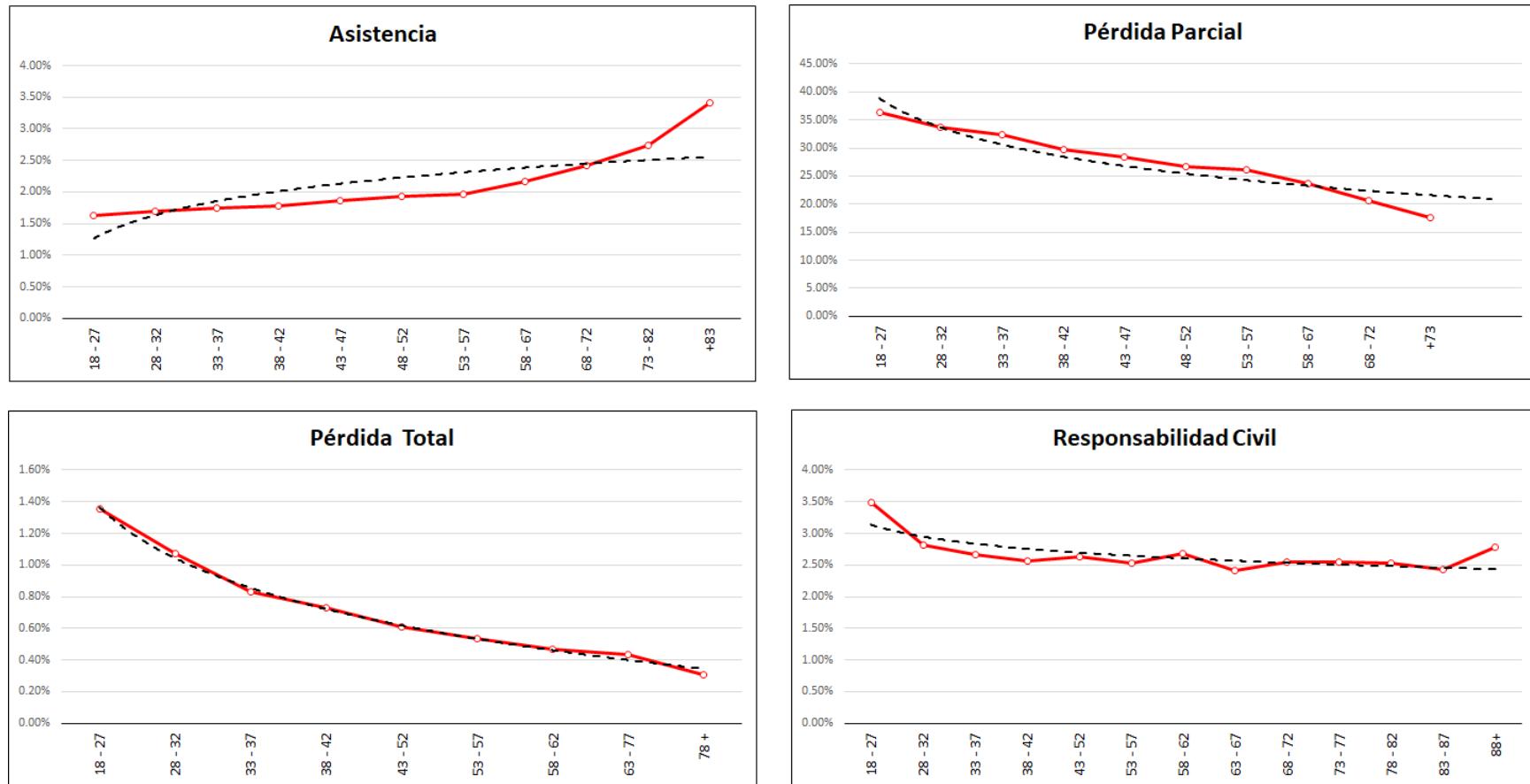
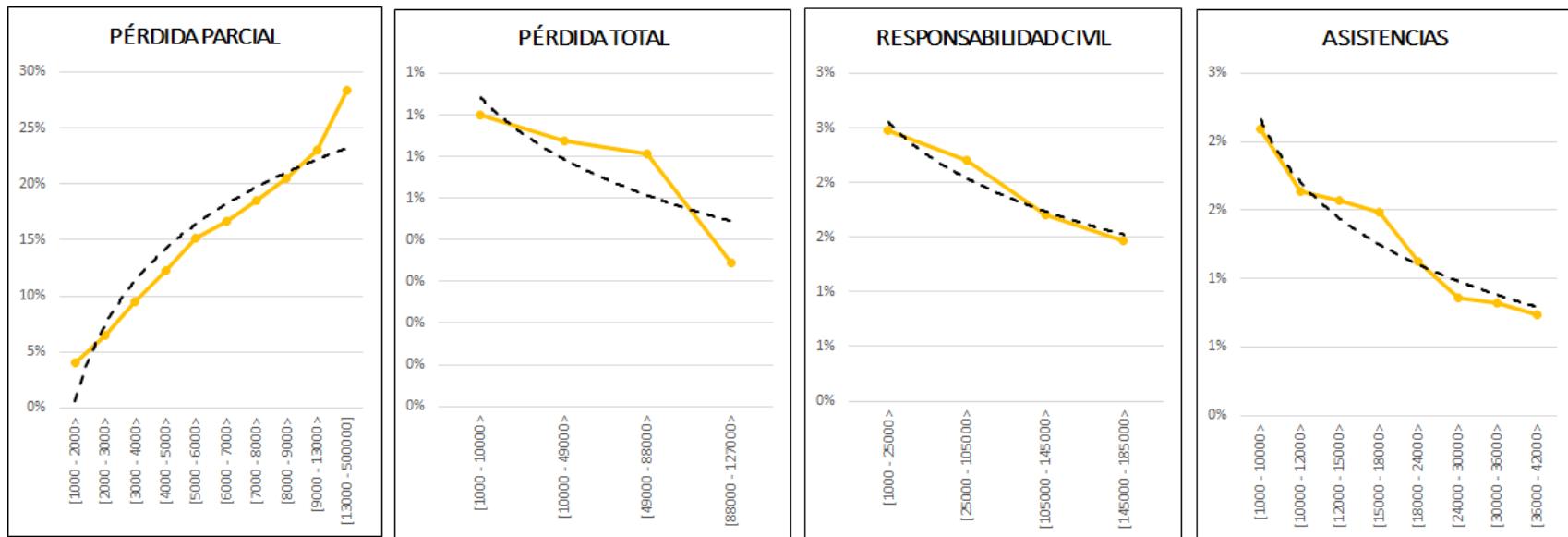


Figure A .15: Edad del asegurado - Agrupado en rangos



Variable: Suma Asegurada

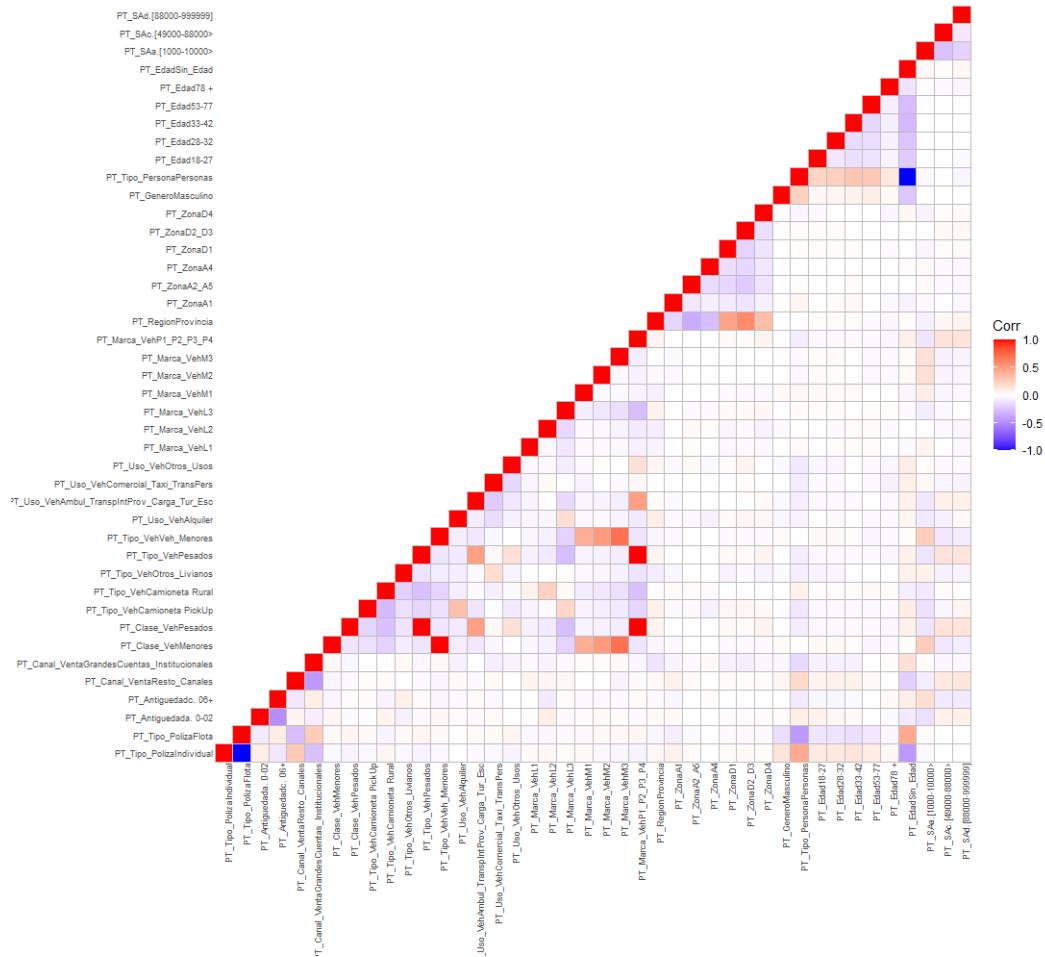
Figure A .16: Rangos de Suma Asegurada del Vehículo



ANEXO B: MODELOS GLM

B.1. Matriz de Correlación entre las variables para detectar posible multicolinealidad

Figure B .17: Matriz de correlación - Pérdida Total

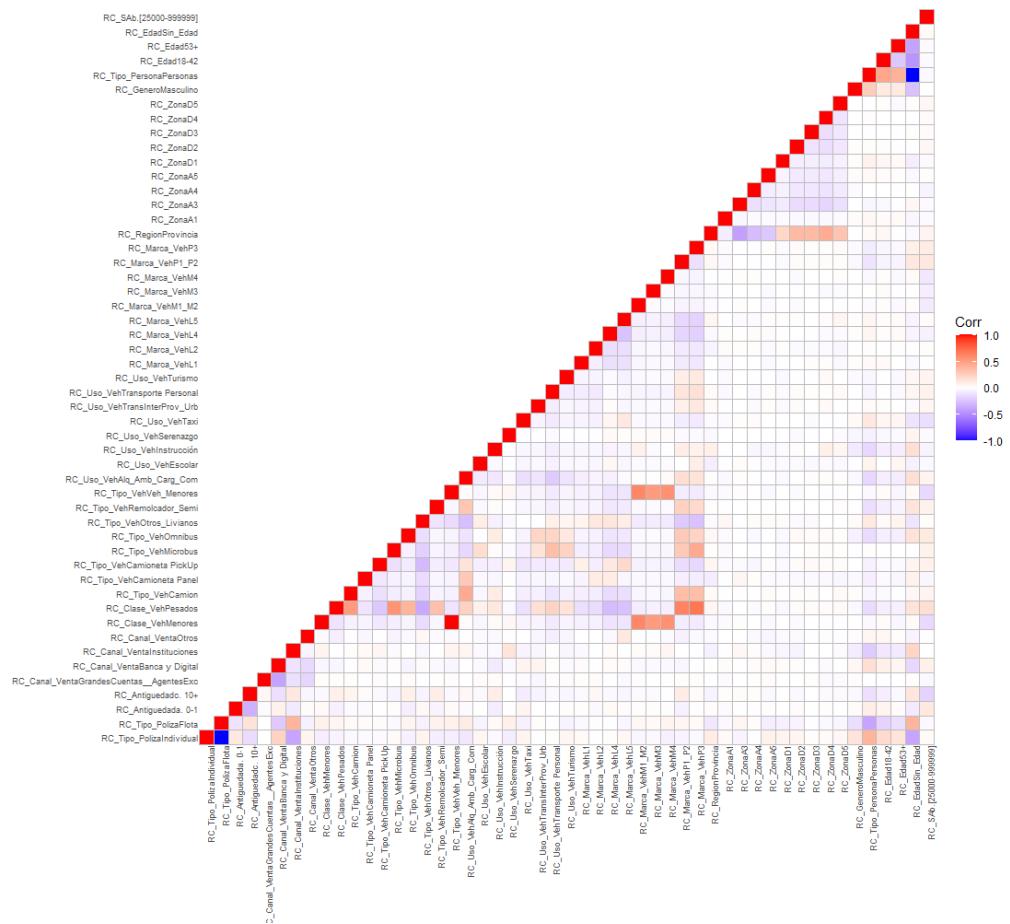


```

1 # MATRIZ DE CORRELACIÓN
2 library(ggplot2)
3 library(ggcorrplot)
4 library(dplyr)
5 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Total/
      glm_Freq.RData")
6 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Total/
      Data_PT.RData")
7 data_PP <- data %>% group_by(PP_Tipo_Poliza, PP_Antiguedad, PP_Canal_Venta, PP_Clase_Veh, PP_Tipo_Veh, PP_Uso_Veh,
     PP_Marca_Veh, PP_Region, PP_Zona, PP_Genero, PP_Tipo_Persona, PP_Edad, PP_SA) %>% summarise(n=n())
8 data_PP <- data_PP[,1:13]
9 model.matrix(~0+, data=data_PP) %>% cor(use="pairwise.complete.obs") %>% ggcorrplot(show.diag=TRUE, type="lower
   ", lab=FALSE) + theme(axis.text.x=element_text(size=7, angle=90, vjust=1, hjust=1, margin=margin(-3,0,0,0)),
   axis.text.y=element_text(size=7, margin=margin(0,-3,0,0)), panel.grid.major=element_blank())

```

Figure B .18: Matriz de correlación - Responsabilidad Civil

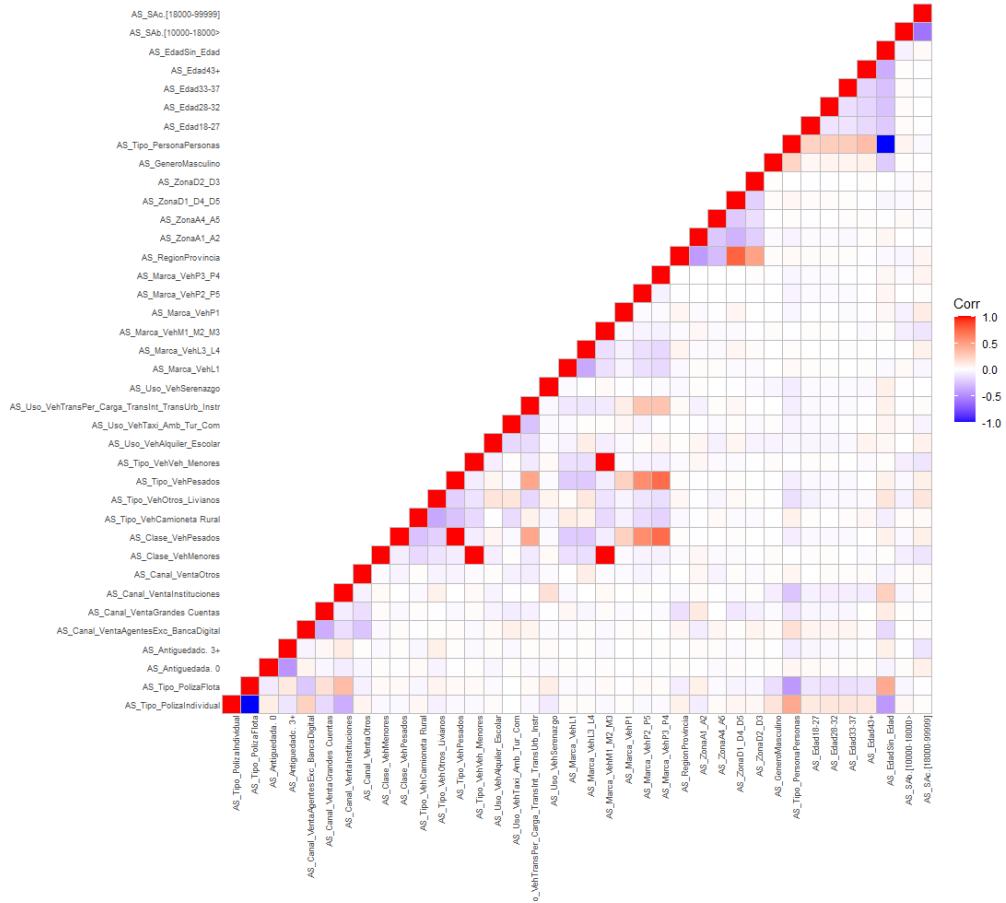


```

1 # MATRIZ DE CORRELACIÓN
2 library(ggplot2)
3 library(ggcorrplot)
4 library(dplyr)
5 load("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Responsabilidad_Civil/glm.Freq.RC.RData")
6 load("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Responsabilidad_Civil/Data_RC.RData")
7 data_RC <- data %>% group_by(RC_Tipo_Poliza ,RC_Antiguedad ,RC_Canal_Venta ,RC_Clase_Veh,RC_Tipo_Veh,RC_Uso_Veh,
8 RC_Marca_Veh,RC_Region ,RC_Zona ,RC_Genero ,RC_Tipo_Persona ,RC_Edad,RC_SA) %>%
9 summarise(n=n())
9 data_RC <- data_RC[,1:13]
10 model.matrix(~0+, data=data_RC) %>% cor(use="pairwise.complete.obs") %>% ggcorrplot(show.diag=TRUE, type="lower",
  lab=FALSE) + theme(axis.text.x=element_text(size=7, angle=90, vjust=1, hjust=1, margin=margin(-3,0,0,0)), axis.text.y=element_text(size=7, margin=margin(0,-3,0,0)), panel.grid.major=element_blank())

```

Figure B .19: Matriz de correlación - Asistencias



```

1 # MATRIZ DE CORRELACIÓN
2 library(ggplot2)
3 library(ggcormplot)
4 library(dplyr)
5 load("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/glm.
       Free.RData")
6 load("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/Data_
       AS.RData")
7 data_AS <- data %>% group_by(AS_Tipo_Poliza, AS_Antiguedad, AS_Canal_Venta, AS_Clase_Veh, AS_Tipo_Veh, AS_Uso_Veh, AS
       _Marca_Veh, AS_Region, AS_Zona, AS_Genero, AS_Tipo_Persona, AS_Edad, AS_SA) %>% summarise(n=n())
8 data_AS <- data_AS[,1:13]
9 model.matrix(~0+, data=data_AS) %>% cor(use="pairwise.complete.obs") %>% ggcormplot(show.diag=TRUE, type="lower
       ", lab=FALSE) + theme(axis.text.x=element_text(size=7, angle=90, vjust=1, hjust=1, margin=margin(-3,0,0,0))
       , axis.text.y=element_text(size=7, margin=margin(0,-3,0,0)), panel.grid.major=element_blank())

```

B.2. Modelos Marginales de la Frecuencia y de la Severidad

Código R: Modelo Frecuencia - Pédida Parcial

```

1 ##### MODELO GLM – FRECUENCIA #####
2 # PAQUETES Y LIBRERÍAS
3 #options(repos = c(cran="http://cran.rstudio.com"))
4 library(RODBC)
5 library(tidyverse) # Incluye dplyr y tidyr para manejar los datos y ggplot para representarlos en gráficos
6 library(plotly) # Para graficos interactivos
7 library(lubridate) # manejar datos y horas
8 library(forcats)
9 library(rchartocolor)
10 library(readr) # para leer y importar archivos
11 library(skimr)
12 library(gcmr)
13 db<-"C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/01_Frecuencia/03_Data/BD_GLM.acedb"
14 con <- odbcConnectAccess2007(db)
15 data <- sqlQuery(con,"select * from PP_Datos_Frecuencia")
16 data.NúmeroPóliza<-format(data.NúmeroPóliza,scientific = FALSE)
17 # Reducimos la data sin el año 2015 y 2020, de acuerdo a los análisis univariados realizados
18 data <- data %>% filter(Año!=2015 & Año!=2020)
19 data <- data %>% filter(PP_Clase_Veh!="Desconocido")
20 # Cambiamos el nombre de algunos niveles de algunas variables
21 data <- data %>% mutate(PP_Antiguedad = case_when(
22   PP_Antiguedad == '0' ~ 'a. 00',PP_Antiguedad == '1' ~ 'b. 01',PP_Antiguedad == '2' ~ 'c. 02',
23   PP_Antiguedad == '3' ~ 'd. 03',PP_Antiguedad == '4' ~ 'e. 04',PP_Antiguedad == '5' ~ 'f. 05',
24   PP_Antiguedad == '6' ~ 'g. 06',PP_Antiguedad == '7' ~ 'h. 07',PP_Antiguedad == '8' ~ 'i. 08',
25   PP_Antiguedad == '9' ~ 'j. 09',PP_Antiguedad == '10' ~ 'k. 10',PP_Antiguedad == '11-13' ~ 'l. 11-13',
26   PP_Antiguedad == '14-19' ~ 'm. 14-19',PP_Antiguedad == '20+' ~ 'n. 20+'
27 ))
28
29 data <- data %>% mutate(PP_SA = case_when(
30   PP_SA == '[1000-2000>' ~ 'a.[1000-2000>',PP_SA == '[2000-3000>' ~ 'b.[2000-3000>',PP_SA == '[3000-4000]' ~ 'c.[3000-4000>',PP_SA == '[4000-5000>' ~ 'd.[4000-5000>',PP_SA == '[5000-6000>' ~ 'e.[5000-6000>',PP_SA == '[6000-7000>' ~ 'f.[6000-7000>',PP_SA == '[7000-8000>' ~ 'g.[7000-8000>',PP_SA == '[8000-9000>' ~ 'h.[8000-9000>',PP_SA == '[9000-13000>' ~ 'i.[9000-13000>',PP_SA == '[13000-99999]' ~ 'j.[13000-99999]'
31 ))
32
33 # Convertimos a factor las variables
34
35 data.Año <- as.factor(data.Año)
36 data.PP_Tipo_Poliza <- as.factor(data.PP_Tipo_Poliza)
37 data.PP_Antiguedad <- as.factor(data.PP_Antiguedad)
38 data.PP_Canal_Venta <- as.factor(data.PP_Canal_Venta)
39 data.PP_Clase_Veh <- as.factor(data.PP_Clase_Veh)
40 data.PP_Uso_Veh <- as.factor(data.PP_Uso_Veh)
41 data.PP_Marca_Veh <- as.factor(data.PP_Marca_Veh)
42 data.PP_Region <- as.factor(data.PP_Region)
43 data.PP_Zona <- as.factor(data.PP_Zona)
44 data.PP_Genero <- as.factor(data.PP_Genero)
45 data.PP_Tipo_Persona <- as.factor(data.PP_Tipo_Persona)
46 data.PP_Edad <- as.factor(data.PP_Edad)
47 data.PP_SA <- as.factor(data.PP_SA)
48
49 # Asignamos el intercepto para cada variable
50 data <- data %>% mutate(
51   Año = fct_relevel(Año, "2022", after = 0),
52   PP_Tipo_Poliza = fct_relevel(PP_Tipo_Poliza , "Individual", after = 0),
53   PP_Antiguedad = fct_relevel(PP_Antiguedad , "c. 02", after = 0),
54   PP_Canal_Venta = fct_relevel(PP_Canal_Venta , "Corredores", after = 0),
55   PP_Clase_Veh = fct_relevel(PP_Clase_Veh , "Livialianos", after = 0),
56   PP_Tipo_Veh = fct_relevel(PP_Tipo_Veh , "Automovil", after = 0),
57   PP_Uso_Veh = fct_relevel(PP_Uso_Veh , "Particular", after = 0),
58   PP_Marca_Veh = fct_relevel(PP_Marca_Veh , "L3", after = 0),
59   PP_Region = fct_relevel(PP_Region , "Lima", after = 0),
60   PP_Zona = fct_relevel(PP_Zona , "A2", after = 0),
61   PP_Genero = fct_relevel(PP_Genero , "Femenino", after = 0),
62   PP_Tipo_Persona = fct_relevel(PP_Tipo_Persona , "Empresa", after = 0 ),
63   PP_Edad = fct_relevel(PP_Edad , "38-42", after = 0),
64   PP_SA = fct_relevel(PP_SA , "j.[13000-99999]", after = 0)
65 )

```

```

70
71 # Verificación de frecuencia por factor
72
73 data %>% group_by(Año) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
74 data %>% group_by(PP_Tipo_Poliza) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
75 data %>% group_by(PP_Antiguedad) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
76 data %>% group_by(PP_Canal_Venta) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
77 data %>% group_by(PP_Clase_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
78 data %>% group_by(PP_Tipo_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
79 data %>% group_by(PP_Uso_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
80 data %>% group_by(PP_Marca_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
81 data %>% group_by(PP_Region) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
82 data %>% group_by(PP_Zona) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
83 data %>% group_by(PP_Genero) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
84 data %>% group_by(PP_Tipo_Persona) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
85 data %>% group_by(PP_Edad) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
86 data %>% group_by(PP_SA) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
87 # Agrupaciones motivadas por modelos previos
88 data <- data %>%
89   mutate(
90     PP_Canal_Venta=fct_collapse(PP_Canal_Venta,"BancaDigital_Otros"=c("Banca y Digital","Otros")),
91     PP_Marca_Veh=fct_collapse(PP_Marca_Veh,"M1_M2_M3"=c("M1","M2","M3")),
92     PP_Zona=fct_collapse(PP_Zona,"A4_A6"=c("A4","A6")))
93 )
94 #####
95 # Agrupamos la data para modelar el GLM
96 #data_model <- data %>% mutate(id=as.numeric(`Número Póliza`)*100+as.numeric(`Número Riesgo`)) %>% group_by(
97   id,Año,PP_Tipo_Poliza,PP_Antiguedad,PP_Canal_Venta,PP_Uso_Veh,PP_Marca_Veh,PP_Zona,PP_Genero,PP_Edad,PP_SA)%>%summarise(Expuesto=sum(Expuesto),Prima_Devengada=sum(Prima_Devengada),Cantidad=sum(Frecuencia),
98   Incurrido=sum(Incurrido))
99 data_model <- data %>%group_by(Año,PP_Tipo_Poliza,PP_Antiguedad,PP_Canal_Venta,PP_Uso_Veh,PP_Marca_Veh,PP_Zona,PP_Genero,PP_Edad,PP_SA)%>%summarise(Expuesto=sum(Expuesto),Prima_Devengada=sum(Prima_Devengada),
100   Cantidad=sum(Frecuencia),Incurrido=sum(Incurrido))
101 # Modelo GLM Pérdida Parcial
102 glm.Frec_PP <- glm(
103   formula = Cantidad ~
104     Año +
105     PP_Tipo_Poliza +
106     PP_Antiguedad +
107     PP_Canal_Venta +
108     #PP_Clase_Veh + (Se correlaciona con la MarcaVeh)
109     #PP_Tipo_Veh + (Se correlaciona con la MarcaVeh)
110     PP_Uso_Veh +
111     PP_Marca_Veh +
112     #PP_Region + (Se correlaciona con la zona)
113     PP_Zona +
114     PP_Genero +
115     #PP_Tipo_Persona + (Se correlaciona con la Edad)
116     PP_Edad +
117     PP_SA +
118     offset(log(Expuesto)),
119   family = poisson(link = "log"),
120   data = data_model
121 )
122 # Desenmascarar posibles multicolinealidades entre las diferentes variables
123 temp <- alias(glm.Frec_PP)
124 temp2 <- as.data.frame(temp.Complete) # Si es vacio, no hay multicolinealidades
125 rm(temp,temp2)
126 # No se realizarán agrupaciones en los niveles de algunas variables por tener un buen modelo
127 # Incluimos las predicciones en la data
128 data <- data %>% mutate(Frecuencia.PRED = glm.Frec_PP.fitted.values)
129 data %>% summarise(sum(Frecuencia), sum(Frecuencia.PRED))
130 # Salvar un data frame con los coeficientes del modelo
131 coefs.Frecuencia_PP <- summary(glm.Frec_PP).coefficients %>%
132   as_tibble() %>%
133   mutate(var_level = glm.Frec_PP.coefficients %>% names()) %>%

```

```

131 select(var_level, everything())
132 # Adicionar manualmente a la tabla los niveles atribuidos al intercept
133 temp <- tibble(
134   var_level = c("Año2022", "PP_Tipo_PolizaIndividual", "PP_Antiguedad_02", "PP_Canal_VentaCorredores", "PP_Uso_
    VehParticular", "PP_Marca_VehL3", "PP_ZonaA2", "PP_GeneroFemenino", "PP_Edad38-42", "PP_SA")[13000-999999]
135 ),
136 Estimate = 0
137 )
138 f.round <- function(x, .digits = 4) round(x, digits = .digits)
139 coefs.Frecuencia.PP <- coefs.Frecuencia.PP %>%
140   bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate)` = exp(Estimate)) %>% mutate_if(is.double,
141   f.round)
142 # Esta función considera los datos de entrenamiento del GLM, agrega la exposición por variable #y la junta a
143 incluir.exposicion <- function(.data, .tbl.coefs){
144   names.vars <- names(.data) %>% select(-Expuesto)
145   n <- length(names.vars)
146   list.temp <- vector(mode = "list", length = n)
147   for(i in seq_along(names.vars)){
148     varname <- sym(names.vars[[i]])
149     list.temp[[i]] <- .data %>%
150       group_by(!varname) %>%
151       summarize(Exposición = f.round(sum(Expuesto), .digits = 0)) %>%
152       ungroup() %>%
153       mutate(var_level = str_c(names.vars[[i]], !!varname)) %>%
154       select(var_level, Exposición)
155   }
156   exp.by.var <- bind_rows(list.temp)
157   return(.tbl.coefs %>% left_join(exp.by.var))
158 }
159 coefs.Frecuencia.PP <- data %>% select(-Frecuencia, -Frecuencia.PRED) %>% incluir.exposicion(coefs.Frecuencia.
160 PP)
161 coefs.Frecuencia.PP %>% DT::datatable()
162 # Guardamos los coeficientes en excel
163 write.csv(coefs.Frecuencia.PP, "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/01_
164 Frecuencia/04_Coeficientes/coefs.Frecuencia.PP.csv", fileEncoding ="Latin1")
165 save(glm.Frec.PP, file = "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_
166 Marginales_Perdida_Parcial/glm.Frec.PP.RData")
167 #data<-data %>% group_by('Número Póliza ', 'Número Riesgo ', 'Año', 'PP_Tipo_Poliza ', 'PP_Antiguedad', 'PP_Canal_Venta', 'PP_
168 Uso_Veh', 'PP_Marca_Veh', 'PP_Zona', 'PP_Genero', 'PP_Edad', 'PP_SA') %>% summarise(Expuesto=sum(Expuesto), Prima_Devengada
169 =sum(Prima_Devengada), Cantidad=sum(Frecuencia), Incurrido=sum(Incurrido))
170 save(data, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
171 Perdida_Parcial/Data_PP.RData")
172 save(data_model, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
173 /Perdida_Parcial/Data_PP.model.RData")

```

Código R: Modelo Severidad - Périda Parcial

```

1 ##### MODELO GLM – SEVERIDAD #####
2 ##### PAQUETES Y LIBRERÍAS #####
3 # PAQUETES Y LIBRERÍAS
4 #options(repos = c(cran="http://cran.rstudio.com"))
5 #install.packages("RODBC")
6 #install.packages("skimr")
7 library(RODBC)
8 library(tidyverse) # Incluye dplyr y tidyr para manejar los datos y ggplot para representarlos en gráficos
9 library(plotly) # Para graficos interactivos
10 library(lubridate) # manejar datas y horas
11 library(forcats)
12 library(rchartcolor)
13 library(readr) # para leer y importar archivos
14 library(skimr)
15 #db2<-"C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/02_CostoMedio/03_Data/BD_GLM_CM.
16 acedb"
17 con2 <- odbcConnectAccess2007(db2)
18 data_CM <- sqlQuery(con2,"select * from PP_Datos_Severidad")
19 # Reducimos la data sin el año 2015 y 2020, de acuerdo a los análisis univariados realizados
20 data_CM <- data_CM %>% filter(Año!=2015 & Año!=2020)
21 data_CM <- data_CM %>% filter(PP_Clase_Veh!="Desconocido")
22 data_CM <- data_CM %>% filter(Incurrido>0) # omitir todas las acciones y saltar al modelo GLM
23 data_CM %>% format(data_CM, Número_Póliza , scientific = FALSE)
24 # Cambiamos el nombre de algunos niveles de algunas variables
25 data_CM <- data_CM %>% mutate(PP_Antiguedad = case_when(
26   PP_Antiguedad == '0' ~ 'a. 00', PP_Antiguedad == '1' ~ 'b. 01', PP_Antiguedad == '2' ~ 'c. 02',
27   PP_Antiguedad == '3' ~ 'd. 03', PP_Antiguedad == '4' ~ 'e. 04', PP_Antiguedad == '5' ~ 'f. 05', PP_Antiguedad == '6' ~ 'g. 06', PP_Antiguedad == '7' ~ 'h. 07', PP_Antiguedad == '8' ~ 'i. 08', PP_Antiguedad == '9' ~ 'j. 09', PP_Antiguedad == '10' ~ 'k. 10', PP_Antiguedad == '11' ~ 'l. 11', PP_Antiguedad == '12' ~ 'm. 12', PP_Antiguedad == '13' ~ 'n. 13', PP_Antiguedad == '14' ~ 'o. 14', PP_Antiguedad == '15' ~ 'p. 15', PP_Antiguedad == '16' ~ 'q. 16', PP_Antiguedad == '17' ~ 'r. 17', PP_Antiguedad == '18' ~ 's. 18', PP_Antiguedad == '19' ~ 't. 19', PP_Antiguedad == '20' ~ 'u. 20', PP_Antiguedad == '21' ~ 'v. 21', PP_Antiguedad == '22' ~ 'w. 22', PP_Antiguedad == '23' ~ 'x. 23', PP_Antiguedad == '24' ~ 'y. 24', PP_Antiguedad == '25' ~ 'z. 25', PP_Antiguedad == '26' ~ 'aa. 26', PP_Antiguedad == '27' ~ 'bb. 27', PP_Antiguedad == '28' ~ 'cc. 28', PP_Antiguedad == '29' ~ 'dd. 29', PP_Antiguedad == '30' ~ 'ee. 30', PP_Antiguedad == '31' ~ 'ff. 31', PP_Antiguedad == '32' ~ 'gg. 32', PP_Antiguedad == '33' ~ 'hh. 33', PP_Antiguedad == '34' ~ 'ii. 34', PP_Antiguedad == '35' ~ 'jj. 35', PP_Antiguedad == '36' ~ 'kk. 36', PP_Antiguedad == '37' ~ 'll. 37', PP_Antiguedad == '38' ~ 'mm. 38', PP_Antiguedad == '39' ~ 'nn. 39', PP_Antiguedad == '40' ~ 'oo. 40', PP_Antiguedad == '41' ~ 'pp. 41', PP_Antiguedad == '42' ~ 'qq. 42', PP_Antiguedad == '43' ~ 'rr. 43', PP_Antiguedad == '44' ~ 'ss. 44', PP_Antiguedad == '45' ~ 'tt. 45', PP_Antiguedad == '46' ~ 'uu. 46', PP_Antiguedad == '47' ~ 'vv. 47', PP_Antiguedad == '48' ~ 'ww. 48', PP_Antiguedad == '49' ~ 'xx. 49', PP_Antiguedad == '50' ~ 'yy. 50', PP_Antiguedad == '51' ~ 'zz. 51', PP_Antiguedad == '52' ~ 'aa. 52', PP_Antiguedad == '53' ~ 'bb. 53', PP_Antiguedad == '54' ~ 'cc. 54', PP_Antiguedad == '55' ~ 'dd. 55', PP_Antiguedad == '56' ~ 'ee. 56', PP_Antiguedad == '57' ~ 'ff. 57', PP_Antiguedad == '58' ~ 'gg. 58', PP_Antiguedad == '59' ~ 'hh. 59', PP_Antiguedad == '60' ~ 'ii. 60', PP_Antiguedad == '61' ~ 'jj. 61', PP_Antiguedad == '62' ~ 'kk. 62', PP_Antiguedad == '63' ~ 'll. 63', PP_Antiguedad == '64' ~ 'mm. 64', PP_Antiguedad == '65' ~ 'nn. 65', PP_Antiguedad == '66' ~ 'oo. 66', PP_Antiguedad == '67' ~ 'pp. 67', PP_Antiguedad == '68' ~ 'qq. 68', PP_Antiguedad == '69' ~ 'rr. 69', PP_Antiguedad == '70' ~ 'ss. 70', PP_Antiguedad == '71' ~ 'tt. 71', PP_Antiguedad == '72' ~ 'uu. 72', PP_Antiguedad == '73' ~ 'vv. 73', PP_Antiguedad == '74' ~ 'ww. 74', PP_Antiguedad == '75' ~ 'xx. 75', PP_Antiguedad == '76' ~ 'yy. 76', PP_Antiguedad == '77' ~ 'zz. 77', PP_Antiguedad == '78' ~ 'aa. 78', PP_Antiguedad == '79' ~ 'bb. 79', PP_Antiguedad == '80' ~ 'cc. 80', PP_Antiguedad == '81' ~ 'dd. 81', PP_Antiguedad == '82' ~ 'ee. 82', PP_Antiguedad == '83' ~ 'ff. 83', PP_Antiguedad == '84' ~ 'gg. 84', PP_Antiguedad == '85' ~ 'hh. 85', PP_Antiguedad == '86' ~ 'ii. 86', PP_Antiguedad == '87' ~ 'jj. 87', PP_Antiguedad == '88' ~ 'kk. 88', PP_Antiguedad == '89' ~ 'll. 89', PP_Antiguedad == '90' ~ 'mm. 90', PP_Antiguedad == '91' ~ 'nn. 91', PP_Antiguedad == '92' ~ 'oo. 92', PP_Antiguedad == '93' ~ 'pp. 93', PP_Antiguedad == '94' ~ 'qq. 94', PP_Antiguedad == '95' ~ 'rr. 95', PP_Antiguedad == '96' ~ 'ss. 96', PP_Antiguedad == '97' ~ 'tt. 97', PP_Antiguedad == '98' ~ 'uu. 98', PP_Antiguedad == '99' ~ 'vv. 99', PP_Antiguedad == '100' ~ 'ww. 100', PP_Antiguedad == '101' ~ 'xx. 101', PP_Antiguedad == '102' ~ 'yy. 102', PP_Antiguedad == '103' ~ 'zz. 103', PP_Antiguedad == '104' ~ 'aa. 104', PP_Antiguedad == '105' ~ 'bb. 105', PP_Antiguedad == '106' ~ 'cc. 106', PP_Antiguedad == '107' ~ 'dd. 107', PP_Antiguedad == '108' ~ 'ee. 108', PP_Antiguedad == '109' ~ 'ff. 109', PP_Antiguedad == '110' ~ 'gg. 110', PP_Antiguedad == '111' ~ 'hh. 111', PP_Antiguedad == '112' ~ 'ii. 112', PP_Antiguedad == '113' ~ 'jj. 113', PP_Antiguedad == '114' ~ 'kk. 114', PP_Antiguedad == '115' ~ 'll. 115', PP_Antiguedad == '116' ~ 'mm. 116', PP_Antiguedad == '117' ~ 'nn. 117', PP_Antiguedad == '118' ~ 'oo. 118', PP_Antiguedad == '119' ~ 'pp. 119', PP_Antiguedad == '120' ~ 'qq. 120', PP_Antiguedad == '121' ~ 'rr. 121', PP_Antiguedad == '122' ~ 'ss. 122', PP_Antiguedad == '123' ~ 'tt. 123', PP_Antiguedad == '124' ~ 'uu. 124', PP_Antiguedad == '125' ~ 'vv. 125', PP_Antiguedad == '126' ~ 'ww. 126', PP_Antiguedad == '127' ~ 'xx. 127', PP_Antiguedad == '128' ~ 'yy. 128', PP_Antiguedad == '129' ~ 'zz. 129', PP_Antiguedad == '130' ~ 'aa. 130', PP_Antiguedad == '131' ~ 'bb. 131', PP_Antiguedad == '132' ~ 'cc. 132', PP_Antiguedad == '133' ~ 'dd. 133', PP_Antiguedad == '134' ~ 'ee. 134', PP_Antiguedad == '135' ~ 'ff. 135', PP_Antiguedad == '136' ~ 'gg. 136', PP_Antiguedad == '137' ~ 'hh. 137', PP_Antiguedad == '138' ~ 'ii. 138', PP_Antiguedad == '139' ~ 'jj. 139', PP_Antiguedad == '140' ~ 'kk. 140', PP_Antiguedad == '141' ~ 'll. 141', PP_Antiguedad == '142' ~ 'mm. 142', PP_Antiguedad == '143' ~ 'nn. 143', PP_Antiguedad == '144' ~ 'oo. 144', PP_Antiguedad == '145' ~ 'pp. 145', PP_Antiguedad == '146' ~ 'qq. 146', PP_Antiguedad == '147' ~ 'rr. 147', PP_Antiguedad == '148' ~ 'ss. 148', PP_Antiguedad == '149' ~ 'tt. 149', PP_Antiguedad == '150' ~ 'uu. 150', PP_Antiguedad == '151' ~ 'vv. 151', PP_Antiguedad == '152' ~ 'ww. 152', PP_Antiguedad == '153' ~ 'xx. 153', PP_Antiguedad == '154' ~ 'yy. 154', PP_Antiguedad == '155' ~ 'zz. 155', PP_Antiguedad == '156' ~ 'aa. 156', PP_Antiguedad == '157' ~ 'bb. 157', PP_Antiguedad == '158' ~ 'cc. 158', PP_Antiguedad == '159' ~ 'dd. 159', PP_Antiguedad == '160' ~ 'ee. 160', PP_Antiguedad == '161' ~ 'ff. 161', PP_Antiguedad == '162' ~ 'gg. 162', PP_Antiguedad == '163' ~ 'hh. 163', PP_Antiguedad == '164' ~ 'ii. 164', PP_Antiguedad == '165' ~ 'jj. 165', PP_Antiguedad == '166' ~ 'kk. 166', PP_Antiguedad == '167' ~ 'll. 167', PP_Antiguedad == '168' ~ 'mm. 168', PP_Antiguedad == '169' ~ 'nn. 169', PP_Antiguedad == '170' ~ 'oo. 170', PP_Antiguedad == '171' ~ 'pp. 171', PP_Antiguedad == '172' ~ 'qq. 172', PP_Antiguedad == '173' ~ 'rr. 173', PP_Antiguedad == '174' ~ 'ss. 174', PP_Antiguedad == '175' ~ 'tt. 175', PP_Antiguedad == '176' ~ 'uu. 176', PP_Antiguedad == '177' ~ 'vv. 177', PP_Antiguedad == '178' ~ 'ww. 178', PP_Antiguedad == '179' ~ 'xx. 179', PP_Antiguedad == '180' ~ 'yy. 180', PP_Antiguedad == '181' ~ 'zz. 181', PP_Antiguedad == '182' ~ 'aa. 182', PP_Antiguedad == '183' ~ 'bb. 183', PP_Antiguedad == '184' ~ 'cc. 184', PP_Antiguedad == '185' ~ 'dd. 185', PP_Antiguedad == '186' ~ 'ee. 186', PP_Antiguedad == '187' ~ 'ff. 187', PP_Antiguedad == '188' ~ 'gg. 188', PP_Antiguedad == '189' ~ 'hh. 189', PP_Antiguedad == '190' ~ 'ii. 190', PP_Antiguedad == '191' ~ 'jj. 191', PP_Antiguedad == '192' ~ 'kk. 192', PP_Antiguedad == '193' ~ 'll. 193', PP_Antiguedad == '194' ~ 'mm. 194', PP_Antiguedad == '195' ~ 'nn. 195', PP_Antiguedad == '196' ~ 'oo. 196', PP_Antiguedad == '197' ~ 'pp. 197', PP_Antiguedad == '198' ~ 'qq. 198', PP_Antiguedad == '199' ~ 'rr. 199', PP_Antiguedad == '200' ~ 'ss. 200', PP_Antiguedad == '201' ~ 'tt. 201', PP_Antiguedad == '202' ~ 'uu. 202', PP_Antiguedad == '203' ~ 'vv. 203', PP_Antiguedad == '204' ~ 'ww. 204', PP_Antiguedad == '205' ~ 'xx. 205', PP_Antiguedad == '206' ~ 'yy. 206', PP_Antiguedad == '207' ~ 'zz. 207', PP_Antiguedad == '208' ~ 'aa. 208', PP_Antiguedad == '209' ~ 'bb. 209', PP_Antiguedad == '210' ~ 'cc. 210', PP_Antiguedad == '211' ~ 'dd. 211', PP_Antiguedad == '212' ~ 'ee. 212', PP_Antiguedad == '213' ~ 'ff. 213', PP_Antiguedad == '214' ~ 'gg. 214', PP_Antiguedad == '215' ~ 'hh. 215', PP_Antiguedad == '216' ~ 'ii. 216', PP_Antiguedad == '217' ~ 'jj. 217', PP_Antiguedad == '218' ~ 'kk. 218', PP_Antiguedad == '219' ~ 'll. 219', PP_Antiguedad == '220' ~ 'mm. 220', PP_Antiguedad == '221' ~ 'nn. 221', PP_Antiguedad == '222' ~ 'oo. 222', PP_Antiguedad == '223' ~ 'pp. 223', PP_Antiguedad == '224' ~ 'qq. 224', PP_Antiguedad == '225' ~ 'rr. 225', PP_Antiguedad == '226' ~ 'ss. 226', PP_Antiguedad == '227' ~ 'tt. 227', PP_Antiguedad == '228' ~ 'uu. 228', PP_Antiguedad == '229' ~ 'vv. 229', PP_Antiguedad == '230' ~ 'ww. 230', PP_Antiguedad == '231' ~ 'xx. 231', PP_Antiguedad == '232' ~ 'yy. 232', PP_Antiguedad == '233' ~ 'zz. 233', PP_Antiguedad == '234' ~ 'aa. 234', PP_Antiguedad == '235' ~ 'bb. 235', PP_Antiguedad == '236' ~ 'cc. 236', PP_Antiguedad == '237' ~ 'dd. 237', PP_Antiguedad == '238' ~ 'ee. 238', PP_Antiguedad == '239' ~ 'ff. 239', PP_Antiguedad == '240' ~ 'gg. 240', PP_Antiguedad == '241' ~ 'hh. 241', PP_Antiguedad == '242' ~ 'ii. 242', PP_Antiguedad == '243' ~ 'jj. 243', PP_Antiguedad == '244' ~ 'kk. 244', PP_Antiguedad == '245' ~ 'll. 245', PP_Antiguedad == '246' ~ 'mm. 246', PP_Antiguedad == '247' ~ 'nn. 247', PP_Antiguedad == '248' ~ 'oo. 248', PP_Antiguedad == '249' ~ 'pp. 249', PP_Antiguedad == '250' ~ 'qq. 250', PP_Antiguedad == '251' ~ 'rr. 251', PP_Antiguedad == '252' ~ 'ss. 252', PP_Antiguedad == '253' ~ 'tt. 253', PP_Antiguedad == '254' ~ 'uu. 254', PP_Antiguedad == '255' ~ 'vv. 255', PP_Antiguedad == '256' ~ 'ww. 256', PP_Antiguedad == '257' ~ 'xx. 257', PP_Antiguedad == '258' ~ 'yy. 258', PP_Antiguedad == '259' ~ 'zz. 259', PP_Antiguedad == '260' ~ 'aa. 260', PP_Antiguedad == '261' ~ 'bb. 261', PP_Antiguedad == '262' ~ 'cc. 262', PP_Antiguedad == '263' ~ 'dd. 263', PP_Antiguedad == '264' ~ 'ee. 264', PP_Antiguedad == '265' ~ 'ff. 265', PP_Antiguedad == '266' ~ 'gg. 266', PP_Antiguedad == '267' ~ 'hh. 267', PP_Antiguedad == '268' ~ 'ii. 268', PP_Antiguedad == '269' ~ 'jj. 269', PP_Antiguedad == '270' ~ 'kk. 270', PP_Antiguedad == '271' ~ 'll. 271', PP_Antiguedad == '272' ~ 'mm. 272', PP_Antiguedad == '273' ~ 'nn. 273', PP_Antiguedad == '274' ~ 'oo. 274', PP_Antiguedad == '275' ~ 'pp. 275', PP_Antiguedad == '276' ~ 'qq. 276', PP_Antiguedad == '277' ~ 'rr. 277', PP_Antiguedad == '278' ~ 'ss. 278', PP_Antiguedad == '279' ~ 'tt. 279', PP_Antiguedad == '280' ~ 'uu. 280', PP_Antiguedad == '281' ~ 'vv. 281', PP_Antiguedad == '282' ~ 'ww. 282', PP_Antiguedad == '283' ~ 'xx. 283', PP_Antiguedad == '284' ~ 'yy. 284', PP_Antiguedad == '285' ~ 'zz. 285', PP_Antiguedad == '286' ~ 'aa. 286', PP_Antiguedad == '287' ~ 'bb. 287', PP_Antiguedad == '288' ~ 'cc. 288', PP_Antiguedad == '289' ~ 'dd. 289', PP_Antiguedad == '290' ~ 'ee. 290', PP_Antiguedad == '291' ~ 'ff. 291', PP_Antiguedad == '292' ~ 'gg. 292', PP_Antiguedad == '293' ~ 'hh. 293', PP_Antiguedad == '294' ~ 'ii. 294', PP_Antiguedad == '295' ~ 'jj. 295', PP_Antiguedad == '296' ~ 'kk. 296', PP_Antiguedad == '297' ~ 'll. 297', PP_Antiguedad == '298' ~ 'mm. 298', PP_Antiguedad == '299' ~ 'nn. 299', PP_Antiguedad == '300' ~ 'oo. 300', PP_Antiguedad == '301' ~ 'pp. 301', PP_Antiguedad == '302' ~ 'qq. 302', PP_Antiguedad == '303' ~ 'rr. 303', PP_Antiguedad == '304' ~ 'ss. 304', PP_Antiguedad == '305' ~ 'tt. 305', PP_Antiguedad == '306' ~ 'uu. 306', PP_Antiguedad == '307' ~ 'vv. 307', PP_Antiguedad == '308' ~ 'ww. 308', PP_Antiguedad == '309' ~ 'xx. 309', PP_Antiguedad == '310' ~ 'yy. 310', PP_Antiguedad == '311' ~ 'zz. 311', PP_Antiguedad == '312' ~ 'aa. 312', PP_Antiguedad == '313' ~ 'bb. 313', PP_Antiguedad == '314' ~ 'cc. 314', PP_Antiguedad == '315' ~ 'dd. 315', PP_Antiguedad == '316' ~ 'ee. 316', PP_Antiguedad == '317' ~ 'ff. 317', PP_Antiguedad == '318' ~ 'gg. 318', PP_Antiguedad == '319' ~ 'hh. 319', PP_Antiguedad == '320' ~ 'ii. 320', PP_Antiguedad == '321' ~ 'jj. 321', PP_Antiguedad == '322' ~ 'kk. 322', PP_Antiguedad == '323' ~ 'll. 323', PP_Antiguedad == '324' ~ 'mm. 324', PP_Antiguedad == '325' ~ 'nn. 325', PP_Antiguedad == '326' ~ 'oo. 326', PP_Antiguedad == '327' ~ 'pp. 327', PP_Antiguedad == '328' ~ 'qq. 328', PP_Antiguedad == '329' ~ 'rr. 329', PP_Antiguedad == '330' ~ 'ss. 330', PP_Antiguedad == '331' ~ 'tt. 331', PP_Antiguedad == '332' ~ 'uu. 332', PP_Antiguedad == '333' ~ 'vv. 333', PP_Antiguedad == '334' ~ 'ww. 334', PP_Antiguedad == '335' ~ 'xx. 335', PP_Antiguedad == '336' ~ 'yy. 336', PP_Antiguedad == '337' ~ 'zz. 337', PP_Antiguedad == '338' ~ 'aa. 338', PP_Antiguedad == '339' ~ 'bb. 339', PP_Antiguedad == '340' ~ 'cc. 340', PP_Antiguedad == '341' ~ 'dd. 341', PP_Antiguedad == '342' ~ 'ee. 342', PP_Antiguedad == '343' ~ 'ff. 343', PP_Antiguedad == '344' ~ 'gg. 344', PP_Antiguedad == '345' ~ 'hh. 345', PP_Antiguedad == '346' ~ 'ii. 346', PP_Antiguedad == '347' ~ 'jj. 347', PP_Antiguedad == '348' ~ 'kk. 348', PP_Antiguedad == '349' ~ 'll. 349', PP_Antiguedad == '350' ~ 'mm. 350', PP_Antiguedad == '351' ~ 'nn. 351', PP_Antiguedad == '352' ~ 'oo. 352', PP_Antiguedad == '353' ~ 'pp. 353', PP_Antiguedad == '354' ~ 'qq. 354', PP_Antiguedad == '355' ~ 'rr. 355', PP_Antiguedad == '356' ~ 'ss. 356', PP_Antiguedad == '357' ~ 'tt. 357', PP_Antiguedad == '358' ~ 'uu. 358', PP_Antiguedad == '359' ~ 'vv. 359', PP_Antiguedad == '360' ~ 'ww. 360', PP_Antiguedad == '361' ~ 'xx. 361', PP_Antiguedad == '362' ~ 'yy. 362', PP_Antiguedad == '363' ~ 'zz. 363', PP_Antiguedad == '364' ~ 'aa. 364', PP_Antiguedad == '365' ~ 'bb. 365', PP_Antiguedad == '366' ~ 'cc. 366', PP_Antiguedad == '367' ~ 'dd. 367', PP_Antiguedad == '368' ~ 'ee. 368', PP_Antiguedad == '369' ~ 'ff. 369', PP_Antiguedad == '370' ~ 'gg. 370', PP_Antiguedad == '371' ~ 'hh. 371', PP_Antiguedad == '372' ~ 'ii. 372', PP_Antiguedad == '373' ~ 'jj. 373', PP_Antiguedad == '374' ~ 'kk. 374', PP_Antiguedad == '375' ~ 'll. 375', PP_Antiguedad == '376' ~ 'mm. 376', PP_Antiguedad == '377' ~ 'nn. 377', PP_Antiguedad == '378' ~ 'oo. 378', PP_Antiguedad == '379' ~ 'pp. 379', PP_Antiguedad == '380' ~ 'qq. 380', PP_Antiguedad == '381' ~ 'rr. 381', PP_Antiguedad == '382' ~ 'ss. 382', PP_Antiguedad == '383' ~ 'tt. 383', PP_Antiguedad == '384' ~ 'uu. 384', PP_Antiguedad == '385' ~ 'vv. 385', PP_Antiguedad == '386' ~ 'ww. 386', PP_Antiguedad == '387' ~ 'xx. 387', PP_Antiguedad == '388' ~ 'yy. 388', PP_Antiguedad == '389' ~ 'zz. 389', PP_Antiguedad == '390' ~ 'aa. 390', PP_Antiguedad == '391' ~ 'bb. 391', PP_Antiguedad == '392' ~ 'cc. 392', PP_Antiguedad == '393' ~ 'dd. 393', PP_Antiguedad == '394' ~ 'ee. 394', PP_Antiguedad == '395' ~ 'ff. 395', PP_Antiguedad == '396' ~ 'gg. 396', PP_Antiguedad == '397' ~ 'hh. 397', PP_Antiguedad == '398' ~ 'ii. 398', PP_Antiguedad == '399' ~ 'jj. 399', PP_Antiguedad == '400' ~ 'kk. 400', PP_Antiguedad == '401' ~ 'll. 401', PP_Antiguedad == '402' ~ 'mm. 402', PP_Antiguedad == '403' ~ 'nn. 403', PP_Antiguedad == '404' ~ 'oo. 404', PP_Antiguedad == '405' ~ 'pp. 405', PP_Antiguedad == '406' ~ 'qq. 406', PP_Antiguedad == '407' ~ 'rr. 407', PP_Antiguedad == '408' ~ 'ss. 408', PP_Antiguedad == '409' ~ 'tt. 409', PP_Antiguedad == '410' ~ 'uu. 410', PP_Antiguedad == '411' ~ 'vv. 411', PP_Antiguedad == '412' ~ 'ww. 412', PP_Antiguedad == '413' ~ 'xx. 413', PP_Antiguedad == '414' ~ 'yy. 414', PP_Antiguedad == '415' ~ 'zz. 415', PP_Antiguedad == '416' ~ 'aa. 416', PP_Antiguedad == '417' ~ 'bb. 417', PP_Antiguedad == '418' ~ 'cc. 418', PP_Antiguedad == '419' ~ 'dd. 419', PP_Antiguedad == '420' ~ 'ee. 420', PP_Antiguedad == '421' ~ 'ff. 421', PP_Antiguedad == '422' ~ 'gg. 422', PP_Antiguedad == '423' ~ 'hh. 423', PP_Antiguedad == '424' ~ 'ii. 424', PP_Antiguedad == '425' ~ 'jj. 425', PP_Antiguedad == '426' ~ 'kk. 426', PP_Antiguedad == '427' ~ 'll. 427', PP_Antiguedad == '428' ~ 'mm. 428', PP_Antiguedad == '429' ~ 'nn. 429', PP_Antiguedad == '430' ~ 'oo. 430', PP_Antiguedad == '431' ~ 'pp. 431', PP_Antiguedad == '432' ~ 'qq. 432', PP_Antiguedad == '433' ~ 'rr. 433', PP_Antiguedad == '434' ~ 'ss. 434', PP_Antiguedad == '435' ~ 'tt. 435', PP_Antiguedad == '436' ~ 'uu. 436', PP_Antiguedad == '437' ~ 'vv. 437', PP_Antiguedad == '438' ~ 'ww. 438', PP_Antiguedad == '439' ~ 'xx. 439', PP_Antiguedad == '440' ~ 'yy. 440', PP_Antiguedad == '441' ~ 'zz. 441', PP_Antiguedad == '442' ~ 'aa. 442', PP_Antiguedad == '443' ~ 'bb. 443', PP_Antiguedad == '444' ~ 'cc. 444', PP_Antiguedad == '445' ~ 'dd. 445', PP_Antiguedad == '446' ~ 'ee. 446', PP_Antiguedad == '447' ~ 'ff. 447', PP_Antiguedad == '448' ~ 'gg. 448', PP_Antiguedad == '449' ~ 'hh. 449', PP_Antiguedad == '450' ~ 'ii. 450', PP_Antiguedad == '451' ~ 'jj. 451', PP_Antiguedad == '452' ~ 'kk. 452', PP_Antiguedad == '453' ~ 'll. 453', PP_Antiguedad == '454' ~ 'mm. 454', PP_Antiguedad == '455' ~ 'nn. 455', PP_Antiguedad == '456' ~ 'oo. 456', PP_Antiguedad == '457' ~ 'pp. 457', PP_Antiguedad == '458' ~ 'qq. 458', PP_Antiguedad == '459' ~ 'rr. 459', PP_Antiguedad == '460' ~ 'ss. 460', PP_Antiguedad == '461' ~ 'tt. 461', PP_Antiguedad == '462' ~ 'uu. 462', PP_Antiguedad == '463' ~ 'vv
```

```

27 PP_Antiguedad == '3' ~ 'd. 03',PP_Antiguedad == '4' ~ 'e. 04',PP_Antiguedad == '5' ~ 'f. 05',
28 PP_Antiguedad == '6' ~ 'g. 06',PP_Antiguedad == '7' ~ 'h. 07',PP_Antiguedad == '8' ~ 'i. 08',
29 PP_Antiguedad == '9' ~ 'j. 09',PP_Antiguedad == '10' ~ 'k. 10',PP_Antiguedad == '11-13' ~ 'l. 11-13',
30 PP_Antiguedad == '14-19' ~ 'm. 14-19',PP_Antiguedad == '20+' ~ 'n. 20+'
31 ))
32
33 data_CM <- data_CM %>% mutate(PP_SA = case_when(
34   PP_SA == '[1000-2000>' ~ 'a.[1000-2000>',PP_SA == '[2000-3000>' ~ 'b.[2000-3000>',PP_SA == '[3000-4000]>' ~
35     'c.[3000-4000>',PP_SA == '[4000-5000>' ~ 'd.[4000-5000>',PP_SA == '[5000-6000>' ~ 'e.[5000-6000>',PP_SA ==
36     '[6000-7000>' ~ 'f.[6000-7000>',PP_SA == '[7000-8000>' ~ 'g.[7000-8000>',PP_SA == '[8000-9000>' ~ 'h
37     '[8000-9000>',PP_SA == '[9000-13000>' ~ 'i.[9000-13000>',PP_SA == '[13000-999999]>' ~ 'j.[13000-999999]>')
38 ))
39 # Convertimos a factor las variables
40 data_CM$Año <- as.factor(data_CM$Año)
41 data_CM$PP_Tipo_Poliza <- as.factor(data_CM$PP_Tipo_Poliza)
42 data_CM$PP_Antiguedad <- as.factor(data_CM$PP_Antiguedad)
43 data_CM$PP_Canal_Venta <- as.factor(data_CM$PP_Canal_Venta)
44 #data_CM$PP_Clase_Veh <- as.factor(data_CM$PP_Clase_Veh)
45 #data_CM$PP_Tipo_Veh <- as.factor(data_CM$PP_Tipo_Veh)
46 data_CM$PP_Uso_Veh <- as.factor(data_CM$PP_Uso_Veh)
47 data_CM$PP_Marca_Veh <- as.factor(data_CM$PP_Marca_Veh)
48 #data_CM$PP_Region <- as.factor(data_CM$PP_Region)
49 data_CM$PP_Zona <- as.factor(data_CM$PP_Zona)
50 data_CM$PP_Genero <- as.factor(data_CM$PP_Genero)
51 data_CM$PP_Tipo_Persona <- as.factor(data_CM$PP_Tipo_Persona)
52 data_CM$PP_Edad <- as.factor(data_CM$PP_Edad)
53 data_CM$PP_SA <- as.factor(data_CM$PP_SA)
54
55 # Asignamos el intercepto para cada variable
56
57 data_CM <- data_CM %>% mutate(
58   Año = fct_relevel(Año, "2022", after = 0),
59   PP_Tipo_Poliza = fct_relevel(PP_Tipo_Poliza, "Individual", after = 0),
60   PP_Antiguedad = fct_relevel(PP_Antiguedad, "c. 02", after = 0),
61   PP_Canal_Venta = fct_relevel(PP_Canal_Venta, "Corredores", after = 0),
62   PP_Clase_Veh = fct_relevel(PP_Clase_Veh, "Livialos", after = 0),
63   PP_Tipo_Veh = fct_relevel(PP_Tipo_Veh, "Automovil", after = 0),
64   PP_Uso_Veh = fct_relevel(PP_Uso_Veh, "Particular", after = 0),
65   PP_Marca_Veh = fct_relevel(PP_Marca_Veh, "L3", after = 0),
66   PP_Region = fct_relevel(PP_Region, "Lima", after = 0),
67   PP_Zona = fct_relevel(PP_Zona, "A2", after = 0),
68   PP_Genero = fct_relevel(PP_Genero, "Femenino", after = 0),
69   PP_Tipo_Persona = fct_relevel(PP_Tipo_Persona, "Empresa", after = 0),
70   PP_Edad = fct_relevel(PP_Edad, "38-42", after = 0),
71   PP_SA = fct_relevel(PP_SA, "j.[13000-999999]", after = 0)
72 )
73
74 # Verificación de frecuencia por factor
75 data_CM %>% group_by(Año) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
76 data_CM %>% group_by(PP_Tipo_Poliza) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
77 data_CM %>% group_by(PP_Antiguedad) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
78 data_CM %>% group_by(PP_Canal_Venta) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
79 data_CM %>% group_by(PP_Clase_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
80 data_CM %>% group_by(PP_Tipo_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
81 data_CM %>% group_by(PP_Uso_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
82 data_CM %>% group_by(PP_Marca_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
83 data_CM %>% group_by(PP_Region) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
84 data_CM %>% group_by(PP_Zona) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
85 data_CM %>% group_by(PP_Genero) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
86 data_CM %>% group_by(PP_Tipo_Persona) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
87 data_CM %>% group_by(PP_Edad) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
88 data_CM %>% group_by(PP_SA) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
89
90 # Agrupaciones motivadas por modelos previos
91 data_CM <- data_CM %>%
92   mutate(
93     PP_Canal_Venta=fct_collapse(PP_Canal_Venta,"BancaDigital_Otros"=c("Banca y Digital","Otros")),
94     PP_Marca_Veh=fct_collapse(PP_Marca_Veh,"M1_M2_M3"=c("M1","M2","M3")),
95     PP_Zona=fct_collapse(PP_Zona,"A4_A6"=c("A4","A6"))
96   )
97
98 # Creamos la variable dependiente
99 data_CM <- data_CM %>% mutate(CM=Incurrido / Cantidad)
100 # Modelo GLM Pérdida Parcial
101 glm.Sev.PP <- glm(

```

```

100 formula = CM ~
101   Año +
102   PP_Tipo_Poliza +
103   PP_Antiguedad +
104   PP_Canal_Venta +
105   #PP_Clase_Veh + (Se correlaciona con la MarcaVeh)
106   #PP_Tipo_Veh + (Se correlaciona con la MarcaVeh)
107   PP_Uso_Veh +
108   PP_Marca_Veh +
109   #PP_Region + (Se correlaciona con la zona)
110   PP_Zona +
111   PP_Genero +
112   #PP_Tipo_Persona + (Se correlaciona con la Edad)
113   PP_Edad +
114   PP_SA,
115   #offset(log(Expuesto)), Se coloca este offset en caso que la variable dependiente hubiera sido solo el
     Incurrido
116 family = Gamma(link = "log"),
117 data = data_CM
118 )
119
120 # Desenmascarar posibles multicolinealidades entre las diferentes variables
121 temp <- alias(glm.Sev.PP)
122 temp2 <- as.data.frame(temp.Complete) # Si es vacio, no hay multicolinealidades
123 rm(temp,temp2)
124 # No se realizarán agrupaciones en los niveles de algunas variables por tener un buen modelo
125 # Incluimos las predicciones en la data
126 data_CM <- data_CM %>% mutate(CM.PRED = glm.Sev.PP.fitted.values)
127 data_CM %>% summarise(sum(Incurrido), sum(CM.PRED))
128 # Salvar un data frame con los coeficientes del modelo
129 coefs.Severidad.PP <- summary(glm.Sev.PP).coefficients %>% as_tibble() %>% mutate(var_level =glm.Sev.PP.
     coefficients %>% names()) %>% select(var_level, everything())
130 # Adicionar manualmente a la tabla los niveles atribuidos al intercept
131 # (y arredondar los valores)
132 temp <- tibble(
133   var_level = c("Año2022", "PP_Tipo_PolizaIndividual", "PP_Antiguedadc_02", "PP_Canal_VentaCorredores", "PP_Uso_
     VehParticular", "PP_Marca_VehL3", "PP_ZonaA2", "PP_GeneroFemenino", "PP_Edad38-42", "PP_SAj.[13000-999999]"),
134   Estimate = 0
135 )
136
137 f.round <- function(x, .digits = 4) round(x, digits = .digits)
138 coefs.Severidad.PP <- coefs.Severidad.PP %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate)` =
     exp(Estimate)) %>% mutate_if(is.double, f.round)
139
140 # Esta función considera los datos de entrenamiento del GLM, agrega la exposición por variable y
141 # la junta a la tabla de coeficientes.
142 incluir.Cantidad <- function(.data, .tbl.coefs){
143   names.vars <- names(.data %>% select(-Cantidad))
144   n <- length(names.vars)
145   list.temp <- vector(mode = "list", length = n)
146   for(i in seq_along(names.vars)){
147     varname <- sym(names.vars[[i]])
148     list.temp[[i]] <- .data %>%
149       group_by(!varname) %>%
150       summarize(Incurrido = f.round(sum(Cantidad), .digits = 0)) %>%
151       ungroup() %>%
152       mutate(var_level = str_c(names.vars[[i]], !varname)) %>%
153       select(var_level, Incurrido)
154   }
155   exp.by.var <- bind_rows(list.temp)
156   return(.tbl.coefs %>% left_join(exp.by.var))
157 }
158 coefs.Severidad.PP <- data_CM %>% select(-CM.PRED) %>% incluir.Cantidad(coefs.Severidad.PP)
159 coefs.Severidad.PP %>% DT::datatable()
160 # Guardamos los coeficientes en excel
161 write.csv(coefs.Severidad.PP,"C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/02_
     CostoMedio/04_Coeficientes/coefs.Severidad.PP.csv",fileEncoding = "Latin1")
162 save(glm.Sev.PP, file = "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_
     Marginales/Perdida_Parcial/glm.Sev.PP.RData")
163 save(data_CM, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
     Perdida_Parcial/Data_PP_model_CM.RData")

```

Código R: Modelo Frecuencia - Pédida Total

```
1 #####
```

```

2 ##### MODELO GLM - FRECUENCIA
3 #####
4
5 # PAQUETES Y LIBRERIAS
6 #options(repos = c(cran="http://cran.rstudio.com"))
7 #install.packages("RODBC")
8 #install.packages("skimr")
9 library(RODBC)
10 library(tidyverse) # Incluye dplyr y tidyverse para manejar los datos y ggplot para representarlos en gráficos
11 library(plotly) # Para gráficos interactivos
12 library(lubridate) # manejar fechas y horas
13 library(forcats)
14 library(rcolorbrewer)
15 library(readr) # para leer y importar archivos
16 library(skimr)
17
18 db<-"/C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/01_Frecuencia/03_Data/BD_GLM.accdb"
19 con <- odbcConnectAccess2007(db)
20 data <- sqlQuery(con,"select * from PT_Datos_Frecuencia")
21 str(data)
22 data.`Número Póliza`<-format(data.`Número Póliza`,scientific = FALSE)
23 # Reducimos la fecha sin el año 2015 y 2020, de acuerdo a los análisis univariados realizados
24 data <- data %>% filter(Año!=2015 & Año!=2020)
25 data <- data %>% filter(PT_Clase_Veh!="Desconocido")
26 # Cambiamos el nombre de algunos niveles de algunas variables
27 data <- data %>% mutate(PT_Antiguedad = case_when(PT_Antiguedad == '0-1' ~ 'a. 0-1',PT_Antiguedad == '2' ~ 'b. 02',PT_Antiguedad == '3-5' ~ 'c. 03-05',PT_Antiguedad == '6' ~ 'd. 06',PT_Antiguedad == '7-16' ~ 'e. 07-16',PT_Antiguedad == '17+' ~ 'f. 17+')
28 ))
29 data <- data %>% mutate(PT_SA = case_when(PT_SA == '[1000-10000>' ~ 'a.[1000-10000]',PT_SA == '[10000-49000>' ~ 'b.[10000-49000]',PT_SA == '[49000-88000>' ~ 'c.[49000-88000]',PT_SA == '[88000-999999]' ~ 'd.[88000-999999]')
30 ))
31 # Convertimos a factor las variables
32 data.Año <- as.factor(data.Año)
33 data.PT_Tipo_Poliza <- as.factor(data.PT_Tipo_Poliza)
34 data.PT_Antiguedad <- as.factor(data.PT_Antiguedad)
35 data.PT_Canal_Venta <- as.factor(data.PT_Canal_Venta)
36 data.PT_Clase_Veh <- as.factor(data.PT_Clase_Veh)
37 data.PT_Tipo_Veh <- as.factor(data.PT_Tipo_Veh)
38 data.PT_Uso_Veh <- as.factor(data.PT_Uso_Veh)
39 data.PT_Marca_Veh <- as.factor(data.PT_Marca_Veh)
40 data.PT_Region <- as.factor(data.PT_Region)
41 data.PT_Zona <- as.factor(data.PT_Zona)
42 data.PT_Genero <- as.factor(data.PT_Genero)
43 data.PT_Tipo_Persona <- as.factor(data.PT_Tipo_Persona)
44 data.PT_Edad <- as.factor(data.PT_Edad)
45 data.PT_SA <- as.factor(data.PT_SA)
46 # Asignamos el intercepto para cada variable
47 data <- data %>% mutate(
48   Año = fct_relevel(Año, "2022", after = 0),
49   PT_Tipo_Poliza = fct_relevel(PT_Tipo_Poliza, "Individual", after = 0),
50   PT_Antiguedad = fct_relevel(PT_Antiguedad, "c. 03-05", after = 0),
51   PT_Canal_Venta = fct_relevel(PT_Canal_Venta, "Corredores", after = 0),
52   PT_Clase_Veh = fct_relevel(PT_Clase_Veh, "Livialianos", after = 0),
53   PT_Tipo_Veh = fct_relevel(PT_Tipo_Veh, "Automovil", after = 0),
54   PT_Uso_Veh = fct_relevel(PT_Uso_Veh, "Particular", after = 0),
55   PT_Marca_Veh = fct_relevel(PT_Marca_Veh, "L4", after = 0),
56   PT_Region = fct_relevel(PT_Region, "Lima", after = 0),
57   PT_Zona = fct_relevel(PT_Zona, "A3", after = 0
58 ),PT_Genero = fct_relevel(PT_Genero, "Femenino", after = 0),
59   PT_Tipo_Persona = fct_relevel(PT_Tipo_Persona, "Empresa", after = 0),
60   PT_Edad = fct_relevel(PT_Edad, "43-52", after = 0),
61   PT_SA = fct_relevel(PT_SA, "b.[10000-49000]", after = 0)
62 )
63 # Verificación de frecuencia por factor
64 data %>% group_by(Año) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
65 data %>% group_by(PT_Tipo_Poliza) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
66 data %>% group_by(PT_Antiguedad) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
67 data %>% group_by(PT_Canal_Venta) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
68 data %>% group_by(PT_Clase_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
69 data %>% group_by(PT_Tipo_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)

```

```

100)
70 data %>% group_by(PT_Uso_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
100)
71 data %>% group_by(PT_Marca_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
100)
72 data %>% group_by(PT_Region) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
100)
73 data %>% group_by(PT_Zona) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
74 data %>% group_by(PT_Genero) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
100)
75 data %>% group_by(PT_Tipo_Persona) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E
,5)*100)
76 data %>% group_by(PT_Edad) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
77 data %>% group_by(PT_SA) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
78
79 # Agrupaciones motivadas por modelos previos
80 data <- data %>%
81   mutate(
82     PT_Antiguedad=fct_collapse(PT_Antiguedad,"a. 0-02"=c("a. 0-1","b. 02"), "c. 06+"=c("d. 06","e. 07-16","f.
17+")),
83     PT_Canal_Venta=fct_collapse(PT_Canal_Venta,"Resto_Canales"=c("Agentes Exclusivos","Banca y Digital","Otros")
,"GrandesCuentas_Institucionales"=c("Grandes Cuentas","Instituciones")),
84     PT_Edad=fct_collapse(PT_Edad,"33-42"=c("33-37","38-42"), "53-77"=c("53-57","58-62","63-77")),
85     PT_Genero=fct_collapse(PT_Genero,"Masculino"=c("Indeterminado","Masculino")),
86     PT_Marca_Veh=fct_collapse(PT_Marca_Veh,"P1_P2_P3_P4"=c("P1","P2","P3","P4")),
87     PT_Uso_Veh=fct_collapse(PT_Uso_Veh,"Ambul_TransIntProv_Carga_Tur_Esc"=c("Ambulancia","Transporte
Interprovincial","Carga","Turismo","Escolar"), "Comercial_Taxi_TransPers"=c("Comercial","Taxi","Transporte
Personal"),"Otros_Uso"=c("Instrucción","Serenazgo","Transporte Urbano")),
88     PT_Zona=fct_collapse(PT_Zona,"A2_A5"=c("A2","A5"), "D2_D3"=c("D2","D3")))
89 )
90 ######
91 # Agrupamos la data para modelar el GLM
92 #data_model <- data %>% mutate(id=as.numeric('Número Póliza')*100+as.numeric('Número Riesgo')) %>% group_by(id
,Año,PP_Tipo_Poliza,PP_Antiguedad,PP_Canal_Venta,PP_Uso_Veh,PP_Marca_Veh,PP_Zona,PP_Genero,PP_Edad,PP_SA)
%>% summarise(Expuesto=sum(Expuesto),Prima_Devengada=sum(Prima_Devengada),Cantidad=sum(Frecuencia),
Incurrido=sum(Incurrido))
93 data_model <- data %>% group_by(Año,
94   PT_Tipo_Poliza ,
95   #PT_Antiguedad ,
96   #PT_Canal_Venta ,
97   PT_Uso_Veh ,
98   PT_Marca_Veh ,
99   PT_Zona ,
100  PT_Genero ,
101  PT_Edad
102  #PP_SA
103   )%>% summarise(Expuesto=sum(Expuesto),Prima_Devengada=sum(Prima_Devengada),
Cantidad=sum(Frecuencia),Incurrido=sum(Incurrido))
104
105
106 # Modelo GLM Pérdida Parcial
107 glm_Frec_PT <- glm(
108   formula = Cantidad ~
109   Año +
110   PT_Tipo_Poliza +
111   #PT_Antiguedad + (No significativa)
112   #PT_Canal_Venta + (No significativa)
113   #PP_Clase_Veh + (Se correlaciona con la MarcaVeh)
114   #PP_Tipo_Veh + (Se correlaciona con la MarcaVeh)
115   PT_Uso_Veh +
116   PT_Marca_Veh +
117   #PP_Region + (Se correlaciona con la zona)
118   PT_Zona +
119   PT_Genero +
120   #PP_Tipo_Persona + (Se correlaciona con la Edad)
121   PT_Edad +
122   #PT_SA + (No significativa)
123   offset(log(Expuesto)),
124   family = poisson(link = "log"),
125   data = data_model
126 )
127 # Desenmascarar posibles multicolinealidades entre las diferentes variables
128 temp <- alias(glm.Frec_PT)
129 temp2 <- as.data.frame(temp.Complete) # Si es vacio, no hay multicolinealidades
130 rm(temp,temp2)
131 # No se realizarán agrupaciones en los niveles de algunas variables por tener un buen modelo

```

```

132 # Incluimos las predicciones en la data
133 data <- data %>% mutate(Frecuencia.PRED = glm.Frec.PT.fitted.values)
134 data %>% summarise(sum(Frecuencia), sum(Frecuencia.PRED))
135 # Salvar un data frame con los coeficientes del modelo
136 coefs.Frecuencia.PT <- summary(glm.Frec.PT).coefficients %>%
137   as_tibble() %>% mutate(var_level = glm.Frec.PT.coefficients %>% names()) %>% select(var_level, everything())
138 # Adicionar manualmente a la tabla los niveles atribuidos al intercept
139 # (y arredondar los valores)
140 temp <- tibble(
141   var_level = c("Año2022", "PT_Tipo_PolizaIndividual", "PT_AntiguedadC_03-05", "PT_Uso_VehParticular", "PT_Marca
142   _VehL4", "PT_ZonaA3", "PT_GeneroFemenino", "PT_Edad43-52"),
143   Estimate = 0
144 )
145 f.round <- function(x, digits = 4) round(x, digits = digits)
146 coefs.Frecuencia.PT <- coefs.Frecuencia.PT %>% bind_rows(temp) %>%
147   arrange(var_level) %>% mutate(`exp(Estimate)` = exp(Estimate)) %>% mutate_if(is.double, f.round)
148 # Esta función considera los datos de entrenamiento del GLM, agrega la exposición por variable # la junta a la
# tabla de coeficientes.
149 incluir.exposicion <- function(.data, .tbl.coefs){
150   names.vars <- names(.data %>% select(-Expuesto))
151   n <- length(names.vars)
152   list.temp <- vector(mode = "list", length = n)
153   for(i in seq_along(names.vars)){
154     varname <- sym(names.vars[[i]])
155     list.temp[[i]] <- .data %>%
156       group_by(!!varname) %>%
157       summarize(Exposición = f.round(sum(Expuesto), digits = 0)) %>%
158       ungroup() %>%
159       mutate(var_level = str_c(names.vars[[i]], !!varname)) %>%
160       select(var_level, Exposición)
161   }
162   exp.by.var <- bind_rows(list.temp)
163   return(.tbl.coefs %>% left_join(exp.by.var))
164 }
165 coefs.Frecuencia.PT <- data %>% select(-Frecuencia, -Frecuencia.PRED) %>% incluir.exposicion(coefs.Frecuencia.
PT)
166 coefs.Frecuencia.PT %>% DT::datatable()
167 # Guardamos los coeficientes en excel
168 write.csv(coefs.Frecuencia.PT, "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/01_
Frecuencia/04_Coeficientes/coefs.Frecuencia.PT.csv", fileEncoding = "Latin1")
169 save(glm.Frec.PT, file = "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_
Marginales/Perdida_Total/glm.Frec.PT.RData")
170 save(data, file = "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
Perdida_Total/Data_PT.RData")
171 save(data_model, file = "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales
/Perdida_Total/Data_PT_model.RData")

```

Código R: Modelo Severidad - Pédida Total

```

1 ##########
2 ##### MODELO GLM – SEVERIDAD #####
3 #####
4 # PAQUETES Y LIBRERIAS
5 #options(repos = c(cran="http://cran.rstudio.com"))
6 #install.packages("RODBC")
7 #install.packages("skimr")
8 library(RODBC)
9 library(tidyverse) # Incluye dplyr y tidyr para manejar los datos y ggplot para representarlos en gráficos
10 library(plotly) # Para graficos interactivos
11 library(lubridate) # manejar datos y horas
12 library(forcats)
13 library(rchartocolor)
14 library(readr) # para leer y importar archivos
15 library(skimr)
16 #db2<- "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/02_CostoMedio/03_Data/BD_GLM_CM.
accdb"
17 #con2 <- odbcConnectAccess2007(db2)
18 #data_CM <- sqlQuery(con2,"select * from PT_Datos_Severidad")
19 #str(data_CM)
20 # Reducimos la data sin el año 2015 y 2020, de acuerdo a los análisis univariados realizados
21 #data_CM <- data_CM %>% filter(Año!=2015 & Año!=2020)
22 #data_CM <- data_CM %>% filter(PT_Clase_Veh!="Desconocido")
23 #data_CM <- data_CM %>% filter(Incurrido>0)
24 #data_CM. `Número Póliza` <- format(data_CM. `Número Póliza`, scientific = FALSE)

```

```

25 data_CM <- data_model %>% filter(Incurrido>0) # omitir todas las acciones y saltar al modelo GLM
26 # Cambiamos el nombre de algunos niveles de algunas variables
27 data_CM <- data_CM %>% mutate(PT_Antiguedad = case_when(
28   PT_Antiguedad == '0-1' ~ 'a. 0-1', PT_Antiguedad == '2' ~ 'b. 02', PT_Antiguedad == '3-5' ~ 'c. 03-05', PT_
29   Antiguedad == '6' ~ 'd. 06', PT_Antiguedad == '7-16' ~ 'e. 07-16', PT_Antiguedad == '17+' ~ 'f. 17+'
30 ))
31 data_CM <- data_CM %>% mutate(PT_SA = case_when(PT_SA == '[1000-10000>' ~ 'a.[1000-10000>', PT_SA == '[10000-49000>' ~ 'b.[10000-49000>', PT_SA == '[49000-88000]' ~ 'c.[49000-88000>', PT_SA == '[88000-999999>' ~ 'd.[88000-999999]')
32 )
33 # Convertimos a factor las variables
34 data_CM$Año <- as.factor(data_CM$Año)
35 data_CM$PT_Tipo_Poliza <- as.factor(data_CM$PT_Tipo_Poliza)
36 data_CM$PT_Antiguedad <- as.factor(data_CM$PT_Antiguedad)
37 data_CM$PT_Canal_Venta <- as.factor(data_CM$PT_Canal_Venta)
38 data_CM$PT_Clase_Veh <- as.factor(data_CM$PT_Clase_Veh)
39 data_CM$PT_Uso_Veh <- as.factor(data_CM$PT_Uso_Veh)
40 data_CM$PT_Marca_Veh <- as.factor(data_CM$PT_Marca_Veh)
41 data_CM$PT_Region <- as.factor(data_CM$PT_Region)
42 data_CM$PT_Zona <- as.factor(data_CM$PT_Zona)
43 data_CM$PT_Genero <- as.factor(data_CM$PT_Genero)
44 data_CM$PT_Tipo_Persona <- as.factor(data_CM$PT_Tipo_Persona)
45 data_CM$PT_Edad <- as.factor(data_CM$PT_Edad)
46 data_CM$PT_SA <- as.factor(data_CM$PT_SA)
47 # Asignamos el intercepto para cada variable
48 data_CM <- data_CM %>% mutate(
49   Año = fct_relevel(Año, "2022", after = 0),
50   PT_Tipo_Poliza = fct_relevel(PT_Tipo_Poliza, "Individual", after = 0),
51   PT_Antiguedad = fct_relevel(PT_Antiguedad, "c. 03-05", after = 0),
52   PT_Canal_Venta = fct_relevel(PT_Canal_Venta, "Corredores", after = 0),
53   PT_Clase_Veh = fct_relevel(PT_Clase_Veh, "Livialos", after = 0),
54   PT_Tipo_Veh = fct_relevel(PT_Tipo_Veh, "Automovil", after = 0), PT_Uso_Veh, "Particular", after = 0),
55   PT_Marca_Veh = fct_relevel(PT_Marca_Veh, "L4", after = 0),
56   PT_Region = fct_relevel(PT_Region, "Lima", after = 0),
57   PT_Zona = fct_relevel(PT_Zona, "A3", after = 0),
58   PT_Genero = fct_relevel(PT_Genero, "Femenino", after = 0),
59   PT_Tipo_Persona = fct_relevel(PT_Tipo_Persona, "Empresa", after = 0),
60   PT_Edad = fct_relevel(PT_Edad, "43-52", after = 0),
61   PT_SA = fct_relevel(PT_SA, "b.[10000-49000>", after = 0)
62 )
63 # Verificación de frecuencia por factor
64 data_CM %>% group_by(Año) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
65 data_CM %>% group_by(PT_Tipo_Poliza) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
66 data_CM %>% group_by(PT_Antiguedad) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
67 data_CM %>% group_by(PT_Canal_Venta) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
68 data_CM %>% group_by(PT_Clase_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
69 data_CM %>% group_by(PT_Uso_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
70 data_CM %>% group_by(PT_Marca_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
71 data_CM %>% group_by(PT_Region) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
72 data_CM %>% group_by(PT_Zona) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
73 data_CM %>% group_by(PT_Genero) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
74 data_CM %>% group_by(PT_Tipo_Persona) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
75 data_CM %>% group_by(PT_Edad) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
76 data_CM %>% group_by(PT_SA) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
77 data_CM %>% group_by(PT_SA) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
78 # Agrupaciones motivadas por modelos previos
79 data_CM <- data_CM %>
80   mutate(
81     PT_Antiguedad=fct_collapse(PT_Antiguedad, "a. 0-02"=c("a. 0-1","b. 02"), "c. 06+"=c("d. 06","e. 07-16","f.
82     17+")),
83     PT_Canal_Venta=fct_collapse(PT_Canal_Venta, "Resto_Canales"=c("Agentes Exclusivos","Banca yDigital","Otros"
84     ), "GrandesCuentas_Institucionales"=c("Grandes Cuentas","Instituciones")),
85     PT_Edad=fct_collapse(PT_Edad, "33-42"=c("33-37","38-42"), "53-77"=c("53-57","58-62","63-77")),
86     PT_Genero=fct_collapse(PT_Genero, "Masculino"=c("Indeterminado","Masculino")),
87     PT_Marca_Veh=fct_collapse(PT_Marca_Veh, "P1_P2_P3_P4"=c("P1","P2","P3","P4")),
88     PT_Uso_Veh=fct_collapse(PT_Uso_Veh, "Ambul_TranspIntProv_Carga_Tur_Esc"=c("Ambulancia","Transporte
89     Interprovincial","Carga","Turismo","Escolar"), "Comercial_Taxi_TransPers"=c("Comercial","Taxi","Transporte
90     Personal"),
91     "Otros_Usos"=c("Instrucción","Serenazgo","Transporte Urbano")
92   ),
93   PT_Zona=fct_collapse(PT_Zona, "A2_A5"=c("A2","A5"), "D2_D3"=c("D2","D3"))
94 )
95 # Creamos la variable dependiente
96 data_CM <- data_CM %>% mutate(CM=Incurrido / Cantidad)
97 # Modelo GLM Pérdida Parcial

```

```

94 glm.Sev.PT <- glm(
95   formula = CM ~
96   Año +
97   PT_Tipo_Poliza +
98   #PT_Antiguedad +
99   #PT_Canal_Venta +
100  #PT_Clase_Veh + (Se correlaciona con la MarcaVeh)
101  #PT_Tipo_Veh + (Se correlaciona con la MarcaVeh)
102  PT_Uso_Veh +
103  PT_Marca_Veh +
104  #PT_Region + (Se correlaciona con la zona)
105  PT_Zona +
106  PT_Genero +
107  #PT_Tipo_Persona + (Se correlaciona con la Edad)
108  PT_Edad,
109  #PP_SA,
110  #offset(log(Expuesto)), Se coloca este offset en caso que la variable dependiente hubiera sido solo el
     Incurrido
111  family = Gamma(link = "log"),
112  data = data_CM
113 )
114 # Desenmascarar posibles multicolinealidades entre las diferentes variables
115 temp <- alias(glm.Sev.PT)
116 temp2 <- as.data.frame(temp.Complete) # Si es vacio, no hay multicolinealidades
117 rm(temp,temp2)
118 # No se realizarán agrupaciones en los niveles de algunas variables por tener un buen modelo
119 # Incluimos las predicciones en la data
120 data_CM <- data_CM %>% mutate(CM.PRED = glm.Sev.PT.fitted.values) data_CM %>% summarise(sum(Incurrido),sum(CM.
     PRED*Cantidad))
121 # Salvar un data frame con los coeficientes del modelo
122 coefs.Severidad.PT <- summary(glm.Sev.PT).coefficients %>as_tibble() %>%mutate(var_level = glm.Sev.PT.
     coefficients %>% names()) %>% select(var_level, everything())
123 # Adicionar manualmente a la tabla los niveles atribuidos al intercept
124 # (y arredondar los valores)
125 temp <- tibble(
126   var_level = c("Año2022","PT_Tipo_PolizaIndividual","#PT_AntiguedadC_03-05","PT_Uso_VehParticular","PT_Marca
     _VehL4","PT_ZonaA3","PT_GeneroFemenino","PT_Edad43-52")
127 ),
128 Estimate = 0
129 )
130 f.round <- function(x, .digits = 4) round(x, digits = .digits)
131 coefs.Severidad.PT <- coefs.Severidad.PT %>%
132   bind_rows(temp) %>%
133   arrange(var_level) %>%
134   mutate(`exp(Estimate)` = exp(Estimate)) %>%
135   mutate_if(is.double, f.round)
136 # Esta función considera los datos de entrenamiento del GLM, agrega la exposición por variable y
137 # la junta a la tabla de coeficientes.
138 incluir.Cantidad <- function(.data, .tbl.coefs){
139   names.vars <- names(.data %>% select(-Cantidad))
140   n <- length(names.vars)
141   list.temp <- vector(mode = "list", length = n)
142   for(i in seq_along(names.vars)){
143     varname <- sym(names.vars[[i]])
144     list.temp[[i]] <- .data %>%
145       group_by(!varname) %>%
146       summarize(Incurrido = f.round(sum(Cantidad), .digits = 0)) %>%
147       ungroup() %>%
148       mutate(var_level = str_c(names.vars[[i]], !varname)) %>%
149       select(var_level, Incurrido)
150   }
151   exp.by.var <- bind_rows(list.temp)
152   return(.tbl.coefs %>% left_join(exp.by.var))
153 }
154 coefs.Severidad.PT <- data_CM %>%
155   select(-CM.PRED) %>%
156   incluir.Cantidad(coefs.Severidad.PT)
157 coefs.Severidad.PT %>% DT:::datatable()
158 # Guardamos los coeficientes en excel
159 write.csv(coefs.Severidad.PT,"C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/02_
     CostoMedio/04_Coeficientes/coefs.Severidad.PT.csv",fileEncoding = "Latin1")
160 save(glm.Sev.PT, file = "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_
     Marginales/Perdida_Total/glm.Sev.PT.RData")
161 save(data_CM, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
     Perdida_Total/Data_PT_model_CM.RData")

```

Código R: Modelo Frecuencia - Responsabilidad Civil

```

1 ##### MODELO GLM – FRECUENCIA #####
2 ##### PAQUETES Y LIBRERÍAS #####
3 # PAQUETES Y LIBRERÍAS
4 #options(repos = c(cran="http://cran.rstudio.com"))
5 library(RODBC)
6 library(tidyverse) # Incluye dplyr y tidyr para manejar los datos y ggplot para representarlos en gráficos
7 library(plotly) # Para graficos interactivos
8 library(lubridate) # manejar datas y horas
9 library(forcats)
10 library(rchartcolor)
11 library(readr) # para leer y importar archivos
12 library(skimr)
13 db<-C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/01_Frecuencia/03_Data/BD_GLM.accdb"
14 con <- odbcConnectAccess2007(db)
15 data <- sqlQuery(con,"select * from RC_Datos_Frecuencia")
16 str(data)
17 data.`Número Poliza`<-format(data.`Número Poliza`,scientific = FALSE)
18 # Reducimos la data sin el año 2015 y 2020, de acuerdo a los análisis univariados realizados
19 data <- data %>% filter(Año!=2015 & Año!=2020)
20 data <- data %>% filter(RC_Clase_Veh!="Desconocido")
21 # Cambiamos el nombre de algunos niveles de algunas variables
22 data <- data %>% mutate(RC_Antiguedad = case_when(RC_Antiguedad == '0-1' ~ 'a. 0-1',RC_Antiguedad == '2-9' ~
23   'b. 2-9',RC_Antiguedad == '10-13' ~ 'c. 10-13',RC_Antiguedad == '14+' ~ 'd. 14+',))
24 )
25 data <- data %>% mutate(RC_SA = case_when(RC_SA == '[1000-25000>' ~ 'a.[1000-25000>',RC_SA == '[25000-105000>' ~
26   'b.[25000-105000>',RC_SA == '[105000-145000]' ~ 'c.[105000-145000]',RC_SA == '[145000-999999]' ~ 'd.
27   [145000-999999]',))
28 # Convertimos a factor las variables
29 data.Año <- as.factor(data.Año)
30 data.RC_Tipo_Poliza <- as.factor(data.RC_Tipo_Poliza)
31 data.RC_Antiguedad <- as.factor(data.RC_Antiguedad)
32 data.RC_Canal_Venta <- as.factor(data.RC_Canal_Venta)
33 data.RC_Clase_Veh <- as.factor(data.RC_Clase_Veh)
34 data.RC_Tipo_Veh <- as.factor(data.RC_Tipo_Veh)
35 data.RC_Uso_Veh <- as.factor(data.RC_Uso_Veh)
36 data.RC_Marca_Veh <- as.factor(data.RC_Marca_Veh)
37 data.RC_Region <- as.factor(data.RC_Region)
38 data.RC_Zona <- as.factor(data.RC_Zona)
39 data.RC_Genero <- as.factor(data.RC_Genero)
40 data.RC_Tipo_Persona <- as.factor(data.RC_Tipo_Persona)
41 data.RC_Edad <- as.factor(data.RC_Edad)
42 data.RC_SA <- as.factor(data.RC_SA)
43 # Asignamos el intercepto para cada variable
44 data <- data %>% mutate(
45   Año = fct_relevel(Año, "2022", after = 0),
46   RC_Tipo_Poliza = fct_relevel(RC_Tipo_Poliza , "Individual", after = 0),
47   RC_Antiguedad = fct_relevel(RC_Antiguedad , "b. 2-9", after = 0),
48   RC_Canal_Venta = fct_relevel(RC_Canal_Venta, "Corredores", after = 0),
49   RC_Clase_Veh = fct_relevel(RC_Clase_Veh, "Livialos", after = 0),
50   RC_Tipo_Veh = fct_relevel(RC_Tipo_Veh, "Automovil", after = 0),
51   RC_Uso_Veh = fct_relevel(RC_Uso_Veh, "Particular", after = 0),
52   RC_Marca_Veh = fct_relevel(RC_Marca_Veh, "L3", after = 0),
53   RC_Region = fct_relevel(RC_Region, "Lima", after = 0),
54   RC_Zona = fct_relevel(RC_Zona, "A2", after = 0),
55   RC_Genero = fct_relevel(RC_Genero, "Femenino", after = 0),
56   RC_Tipo_Persona = fct_relevel(RC_Tipo_Persona, "Empresa", after = 0),
57   RC_Edad = fct_relevel(RC_Edad, "43-52", after = 0),
58   RC_SA = fct_relevel(RC_SA, "a.[1000-25000]>", after = 0)
59 )
60 # Verificación de frecuencia por factor
61 data %>% group_by(Año) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
62 data %>% group_by(RC_Tipo_Poliza) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>%mutate(Freq=round(C/E
63   ,5)*100)
64 data %>% group_by(RC_Antiguedad) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>%mutate(Freq=round(C/E
65   ,5)*100)
66 data %>% group_by(RC_Canal_Venta) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>%mutate(Freq=round(C/E
67   ,5)*100)
68 data %>% group_by(RC_Clase_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>%mutate(Freq=round(C/E,5)*
69   100)
70 data %>% group_by(RC_Tipo_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
71   100)

```

```

67 data %>% group_by(RC_Uso_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
68   100)
69 data %>% group_by(RC_Marca_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
70   100)
71 data %>% group_by(RC_Region) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
72   100)
73 data %>% group_by(RC_Zona) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
74 data %>% group_by(RC_Genero) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
75   100)
76 data %>% group_by(RC_Tipo_Persona) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E
77   ,5)*100)
78 data %>% group_by(RC_Edad) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
79 data %>% group_by(RC_SA) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
80 # Agrupaciones motivadas por modelos previos
81 data <- data %>%
82   mutate(
83     RC_Antiguedad=fct_collapse(RC_Antiguedad,"c. 10+"=c("c. 10-13","d. 14+")),
84     RC_Canal_Venta=fct_collapse(RC_Canal_Venta,"GrandesCuentas__AgentesExc=c("Agentes Exclusivos","Grandes
85       Cuentas")),
86     RC_Edad=fct_collapse(RC_Edad,"18-42"=c("18-27","28-32","33-37","38-42"), "53+"=c("53-57","58-62","63-67",
87       "73-77","78-82","83-87","88+")),
88     RC_Genero=fct_collapse(RC_Genero,"Masculino"=c("Indeterminado","Masculino")),
89     RC_Marca_Veh=fct_collapse(RC_Marca_Veh,"MI_M2"=c("MI","M2"), "P1_P2"=c("P1","P2")),
90     RC_SA=fct_collapse(RC_SA,"b.[25000-999999]"=c("b.[25000-105000>","c.[105000-145000>","d.[145000-999999]"))
91     ,
92     RC_Uso_Veh=fct_collapse(RC_Uso_Veh,"Alq_Amb_Carg_Com"=c("Alquiler","Ambulancia","Carga","Comercial"),
93       "TransInterProv_Urb"=c("Transporte Interprovincial","Transporte Urbano"))
94   )
95 #####
96 # Agrupamos la data para modelar el GLM
97 data_model <- data %>% mutate(id=as.numeric(`Número Póliza`)*100+as.numeric(`Número Riesgo`)) %>% group_by(id
98   ,Año,PP_Tipo_Poliza,PP_Antiguedad,PP_Canal_Venta,PP_Uso_Veh,PP_Marca_Veh,PP_Zona,PP_Genero,PP_Edad,PP_SA
99   )%>%summarise(Expuesto=sum(Expuesto),Prima_Devengada=sum(Prima_Devengada),Cantidad=sum(Frecuencia),
100   Incurrido=sum(Incurrido))
101 data_model <- data %>% group_by(Año,
102   #PT_Tipo_Poliza,
103   RC_Antiguedad,
104   RC_Canal_Venta,
105   RC_Uso_Veh,
106   RC_Marca_Veh,
107   RC_Zona,
108   #PT_Genero,
109   RC_Edad
110   #PP_SA
111 )%>% summarise(Expuesto=sum(Expuesto),Prima_Devengada=sum(Prima_Devengada),Cantidad=sum(Frecuencia),Incurrido=
112   sum(Incurrido))
113 # Modelo GLM Pérdida Parcial
114 glm.Frec.RC <- glm(
115   formula = Cantidad ~
116     Año +
117     #RC_Tipo_Poliza + (no significativo)
118     RC_Antiguedad + #(No significativa)
119     RC_Canal_Venta + #(No significativa)
120     #RC_Clase_Veh + (Se correlaciona con la MarcaVeh)
121     #RC_Tipo_Veh + (Se correlaciona con la MarcaVeh)
122     RC_Uso_Veh +
123     RC_Marca_Veh +
124     #RC_Region + (Se correlaciona con la zona)
125     RC_Zona +
126     #RC_Genero + (no significativo)
127     #RC_Tipo_Persona + (Se correlaciona con la Edad)
128     RC_Edad +
129     #RC_SA + #(No significativa)
130     offset(log(Expuesto)),
131     family = poisson(link = "log"),
132     data = data_model
133   )
134 # Desenmascarar posibles multicolinealidades entre las diferentes variables
135 temp <- alias(glm.Frec.RC)
136 temp2 <- as.data.frame(temp.Complete) # Si es vacio, no hay multicolinealidades
137 rm(temp,temp2)
138 # No se realizarán agrupaciones en los niveles de algunas variables por tener un buen modelo
139 # Incluimos las predicciones en la data
140 data <- data %>% mutate(Frecuencia.PRED = glm.Frec.RC.fitted.values)

```

```

131 data %>% summarise(sum(Frecuencia), sum(Frecuencia.PRED))
132 # Salvar un data frame con los coeficientes del modelo
133 coefs.Frecuencia.RC <- summary(glm.Frec.RC).coefficients %>%as_tibble() %>%mutate(var_level =glm.Frec.RC.
134   coefficients %>% names()) %>% select(var_level, everything())
135 # Adicionar manualmente a la tabla los niveles atribuidos al intercept
136 # (y arredondar los valores)
137 temp <- tibble(
138   var_level = c("Año2022", "RC_Antiguedadb_ 2-9", "RC_Canal_VentaCorredores", "RC_Uso_VehParticular", "RC_Marca_
139   VehL3", "RC_ZonaA2", "RC_Edad43-52")
140   ),
141   Estimate = 0
140 )
141 f.round <- function(x, .digits = 4) round(x, digits = .digits)
142 coefs.Frecuencia.RC <- coefs.Frecuencia.RC %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate
143   )` = exp(Estimate)) %>% mutate_if(is.double, f.round)
144 # Esta función considera los datos de entrenamiento del GLM, agrega la exposición por variable y
145 # la junta a la tabla de coeficientes.
146 incluir.exposición <- function(.data, .tbl.coefs){
147   names.vars <- names(.data %>% select(-Expuesto))
148   n <- length(names.vars)
149   list.temp <- vector(mode = "list", length = n)
150   for(i in seq_along(names.vars)){
151     varname <- sym(names.vars[[i]])
152     list.temp[[i]] <- .data %>%
153       group_by(!varname) %>%
154       summarize(Exposición = f.round(sum(Expuesto), .digits = 0)) %>%
155       ungroup() %>%
156       mutate(var_level = str_c(names.vars[[i]], !varname)) %>%
157       select(var_level, Exposición)
158   }
159   exp.by.var <- bind_rows(list.temp)
160   return(.tbl.coefs %>% left_join(exp.by.var))
161 }
162 coefs.Frecuencia.RC <- data %>%
163   select(-Frecuencia, -Frecuencia.PRED) %>%incluir.exposición(coefs.Frecuencia.RC)
164 coefs.Frecuencia.RC %>% DT::datatable()
165 # Guardamos los coeficientes en excel
166 write.csv(coefs.Frecuencia.RC, "C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/01_
167 Frecuencia/04_Coeficientes/coefs.Frecuencia.RC.csv", fileEncoding ="Latin1")
167 save(glm.Frec.RC, file = "C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_
168 Marginales/Responsabilidad_Civil/glm.Frec.RC.RData")
168 save(data, file="C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
169 Responsabilidad_Civil/Data_RC.RData")
169 save(data_model, file="C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales_
/Responsabilidad_Civil/Data_RC_model.RData")

```

Código R: Modelo severidad - Responsabilidad Civil

```

1 ##### MODELO GLM – SEVERIDAD #####
2 ##### PAQUETES Y LIBRERÍAS
3 #options(repos = c(cran="http://cran.rstudio.com"))
4 #install.packages("RODBC")
5 #install.packages("skimr")
6 library(RODBC)
7 library(tidyverse) # Incluye dplyr y tidyr para manejar los datos y ggplot para representarlos en gráficos
8 library(plotly) # Para graficos interactivos
9 library(lubridate) # manejar datas y horas
10 library(forcats)
11 library(rchartcolor)
12 library(readr) # para leer y importar archivos
13 library(skimr)
14 #db2<-dbConnect("ODBC", "BD_GL_M")
15 #db2<-dbConnect("ODBC", "BD_GL_CM")
16 #db2<-dbConnect("ODBC", "BD_GL_CM")
17 #con2 <- db2
18 #data_CM <- sqlQuery(con2,"select * from RC_Datos_Severidad")
19 #str(data_CM)
20 # Reducimos la data sin el año 2015 y 2020, de acuerdo a los análisis univariados realizados
21 #data_CM <- data_CM %>% filter(Año!=2015 & Año!=2020)
22 #data_CM <- data_CM %>% filter(RC_Clase_Veh!="Desconocido")
23 #data_CM <- data_CM %>% filter(Incurrido>0)
24 #data_CM.‘Número Póliza’<-format(data_CM.‘Número Póliza’,scientific = FALSE)
25 data_CM <- data_CM %>% filter(Incurrido>0) # omitir todas las acciones y saltar al modelo GLM

```

```

26 # Cambiamos el nombre de algunos niveles de algunas variables
27 data_CM <- data_CM %>% mutate(RC_Antiguedad = case_when(RC_Antiguedad == '0-1' ~ 'a. 0-1', RC_Antiguedad == '2-9' ~ 'b. 2-9', RC_Antiguedad == '10-13' ~ 'c. 10-13', RC_Antiguedad == '14+' ~ 'd. 14+')
28 ))
29 data_CM <- data_CM %>% mutate(RC_SA = case_when(RC_SA == '[1000-25000>' ~ 'a.[1000-25000>', RC_SA == '[25000-105000>' ~ 'b.[25000-105000>', RC_SA == '[105000-145000>' ~ 'c.[105000-145000>', RC_SA == '[145000-999999]' ~ 'd.[145000-999999]')
30 ))
31 # Convertimos a factor las variables
32 data_CM$Año <- as.factor(data_CM$Año)
33 data_CM$RC_Tipo_Poliza <- as.factor(data_CM$RC_Tipo_Poliza)
34 data_CM$RC_Antiguedad <- as.factor(data_CM$RC_Antiguedad)
35 data_CM$RC_Canal_Venta <- as.factor(data_CM$RC_Canal_Venta)
36 data_CM$RC_Clase_Veh <- as.factor(data_CM$RC_Clase_Veh)
37 data_CM$RC_Tipo_Veh <- as.factor(data_CM$RC_Tipo_Veh)
38 data_CM$RC_Uso_Veh <- as.factor(data_CM$RC_Uso_Veh)
39 data_CM$RC_Marca_Veh <- as.factor(data_CM$RC_Marca_Veh)
40 data_CM$RC_Region <- as.factor(data_CM$RC_Region)
41 data_CM$RC_Zona <- as.factor(data_CM$RC_Zona)
42 data_CM$RC_Genero <- as.factor(data_CM$RC_Genero)
43 data_CM$RC_Tipo_Persona <- as.factor(data_CM$RC_Tipo_Persona)
44 data_CM$RC_Edad <- as.factor(data_CM$RC_Edad)
45 data_CM$RC_SA <- as.factor(data_CM$RC_SA)
46 # Asignamos el intercepto para cada variable
47 data_CM <- data_CM %>% mutate(
48   Año = fct_relevel(Año, "2022", after = 0),
49   RC_Tipo_Poliza = fct_relevel(RC_Tipo_Poliza, "Individual", after = 0),
50   RC_Antiguedad = fct_relevel(RC_Antiguedad, "b. 2-9", after = 0),
51   RC_Canal_Venta = fct_relevel(RC_Canal_Venta, "Corredores", after = 0),
52   RC_Clase_Veh = fct_relevel(RC_Clase_Veh, "Livistano", after = 0),
53   RC_Tipo_Veh = fct_relevel(RC_Tipo_Veh, "Automovil", after = 0),
54   RC_Uso_Veh = fct_relevel(RC_Uso_Veh, "Particular", after = 0),
55   RC_Marca_Veh = fct_relevel(RC_Marca_Veh, "L3", after = 0),
56   RC_Region = fct_relevel(RC_Region, "Lima", after = 0),
57   RC_Zona = fct_relevel(RC_Zona, "A2", after = 0),
58   RC_Genero = fct_relevel(RC_Genero, "Femenino", after = 0),
59   RC_Tipo_Persona = fct_relevel(RC_Tipo_Persona, "Empresa", after = 0),
60   RC_Edad = fct_relevel(RC_Edad, "43-52", after = 0),
61   RC_SA = fct_relevel(RC_SA, "a.[1000-25000]", after = 0)
62 )
63 # Verificación de frecuencia por factor
64 data_CM %>% group_by(Año) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
65 data_CM %>% group_by(RC_Tipo_Poliza) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
66 data_CM %>% group_by(RC_Antiguedad) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
67 data_CM %>% group_by(RC_Canal_Venta) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
68 data_CM %>% group_by(RC_Clase_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
69 data_CM %>% group_by(RC_Tipo_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
70 data_CM %>% group_by(RC_Uso_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
71 data_CM %>% group_by(RC_Marca_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
72 data_CM %>% group_by(RC_Region) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
73 data_CM %>% group_by(RC_Zona) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
74 data_CM %>% group_by(RC_Genero) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
75 data_CM %>% group_by(RC_Tipo_Persona) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
76 data_CM %>% group_by(RC_Edad) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
77 data_CM %>% group_by(RC_SA) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
78 # Agrupaciones motivadas por modelos previos
79 data_CM <- data_CM %>%
80   mutate(
81     RC_Antiguedad=fct_collapse(RC_Antiguedad,"c. 10+"=c("c. 10-13","d. 14+")),
82     RC_Canal_Venta=fct_collapse(RC_Canal_Venta,"GrandesCuentas__AgentesExc=c("Agentes Exclusivos","Grandes
83     Cuentas"))),
84     RC_Edad=fct_collapse(RC_Edad,"18-42=c("18-27","28-32","33-37","38-42"),"53+"=c("53-57","58-62","63-67",
85     "73-77","68-72","73-77","78-82","83-87","88+")),
86     RC_Genero=fct_collapse(RC_Genero,"Masculino=c("Indeterminado","Masculino")),
87     RC_Marca_Veh=fct_collapse(RC_Marca_Veh,"M1_M2=c("M1","M2"),"P1_P2=c("P1","P2")),
88     RC_SA=fct_collapse(RC_SA,"b.[25000-999999]"=c("b.[25000-105000]>","c.[105000-145000]>","d.[145000-999999]")),
89     RC_Uso_Veh=fct_collapse(RC_Uso_Veh,"Alq_Amb_Carg_Com=c("Alquiler","Ambulancia","Carga","Comercial"),
90     "TransInterProv_Urb=c("Transporte Interprovincial","Transporte Urbano"))
91   )
92 # Creamos las variable dependiente
93 data_CM <- data_CM %>% mutate(CM=Incurrido / Cantidad)
94 # Modelo GLM Pérdida Parcial
95 glm.Sev.RC <- glm(
96   formula = CM ~
97     Año +

```

```

95  #RC_Tipo_Poliza +
96  RC_Antiguedad +
97  RC_Canal_Venta +
98  #RC_Clase_Veh + (Se correlaciona con la MarcaVeh)
99  #RC_Tipo_Veh + (Se correlaciona con la MarcaVeh)
100 RC_Uso_Veh +
101 RC_Marca_Veh +
102 #RC_Region + (Se correlaciona con la zona)
103 RC_Zona +
104 #RC_Genero +
105 #RC_Tipo_Persona + (Se correlaciona con la Edad)
106 RC_Edad,
107 #RC_SA,
108 #offset(log(Expuesto)), Se coloca este offset en caso que la variable dependiente hubiera sido solo el
109   Incurrido
110 family = Gamma(link = "log"),
111 data = data_CM
111 )
112 # Desenmascarar posibles multicolinealidades entre las diferentes variables
113 temp <- alias(glm.Sev.RC)
114 temp2 <- as.data.frame(temp.Complete) # Si es vacio, no hay multicolinealidades
115 rm(temp,temp2)
116 # No se realizarán agrupaciones en los niveles de algunas variables por tener un buen modelo
117 # Incluimos las predicciones en la data
118 data_CM <- data_CM %>% mutate(CM.PRED = glm.Sev.RC.fitted.values)
119 data_CM %>% summarise(sum(Incurrido), sum(CM.PRED*Cantidad))
120 # Salvar un data frame con los coeficientes del modelo
121 coefs.Severidad.RC <- summary(glm.Sev.RC).coefficients %>%
122   as_tibble() %>%
123   mutate(var_level = glm.Sev.RC.coefficients %>% names()) %>%
124   select(var_level, everything())
125 # Adicionar manualmente a la tabla los niveles atribuidos al intercept
126 # (y arredondar los valores)
127 temp <- tibble(
128   var_level = c("Año2022","RC_Antiguedadb_ 2-9","RC_Canal_VentaCorredores","RC_Uso_VehParticular","RC_Marca_
129   VehL3","RC_ZonaA2","RC_Edad43-52"),
130   Estimate = 0
130 )
131 f.round <- function(x, .digits = 4) round(x, digits = .digits)
132 coefs.Severidad.RC <- coefs.Severidad.RC %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate)` =
133   = exp(Estimate)) %>% mutate_if(is.double, f.round)
134 # Esta función considera los datos de entrenamiento del GLM, agrega la exposición por variable y
135 # la junta a la tabla de coeficientes.
136 incluir.Cantidad <- function(.data, .tbl.coefs){
137   names.vars <- names(.data %>% select(-Cantidad))
138   n <- length(names.vars)
139   list.temp <- vector(mode = "list", length = n)
140   for(i in seq_along(names.vars)){
141     varname <- sym(names.vars[[i]])
142     list.temp[[i]] <- .data %>%
143       group_by(!varname) %>%
144       summarize(Incurrido = f.round(sum(Cantidad), .digits = 0)) %>%
145       ungroup() %>%
146       mutate(var_level = str_c(names.vars[[i]], !varname)) %>%
147       select(var_level, Incurrido)
148   }
149   exp.by.var <- bind_rows(list.temp)
150   return(.tbl.coefs %>% left_join(exp.by.var))
150 }
151 coefs.Severidad.RC <- data_CM %>%
152   select(-CM.PRED) %>%
153   incluir.Cantidad(coefs.Severidad.RC)
154 coefs.Severidad.RC %>% DT:::datatable()
155 # Guardamos los coeficientes en excel
156 write.csv(coefs.Severidad.RC,"C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/02_
157 CostoMedio/04_Coeficientes/coefs.Severidad.RC.csv",fileEncoding = "Latin1")
157 save(glm.Sev.RC, file = "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_
158 Marginales/Responsabilidad_Civil/glm.Sev.RC.RData")
158 save(data_CM, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
158 Responsabilidad_Civil/Data_RC_model_CM.RData")
```

Código R: Modelo Frecuencia - Asistencias

```

1 #####
2 ##### MODELO GLM – FRECUENCIA
```

```

3 ######
4 # PAQUETES Y LIBRERIAS
5 #options(repos = c(cran="http://cran.rstudio.com"))
6 #install.packages("RODBC")
7 #install.packages("skimr")
8 library(RODBC)
9 library(tidyverse) # Incluye dplyr y tidyverse para manejar los datos y ggplot para representarlos en gráficos
10 library(plotly) # Para graficos interactivos
11 library(lubridate) # manejar datas y horas
12 library(forcats)
13 library(rchartcolor)
14 library(readr) # para leer y importar archivos
15 library(skimr)
16 db<-"C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/01_Frecuencia/03_Data/BD_GLM.accdb"
17 con <- odbcConnectAccess2007(db)
18 data <- sqlQuery(con,"select * from AS_Datos_Frecuencia")
19 str(data)
20 data.Número_Póliza <- format(data.Número_Póliza ,scientific = FALSE)
21 # Reducimos la data sin el año 2015 y 2020, de acuerdo a los análisis univariados realizados
22 data <- data %>% filter(Año!=2015 & Año!=2020)
23 data <- data %>% filter(AS_Clase_Veh!="Desconocido")
24 data <- data %>% filter(Frecuencia<100)
25 # Cambiamos el nombre de algunos niveles de algunas variables
26 data <- data %>% mutate(AS_Antiguedad = case_when(
27   AS_Antiguedad == '0' ~ 'a. 0',
28   AS_Antiguedad == '1-2' ~ 'b. 1-2',
29   AS_Antiguedad == '3' ~ 'c. 3',
30   AS_Antiguedad == '4-5' ~ 'd. 4-5',
31   AS_Antiguedad == '6' ~ 'e. 6',
32   AS_Antiguedad == '7' ~ 'f. 7',
33   AS_Antiguedad == '8+' ~ 'g. 8+'
34 ))
35 data <- data %>% mutate(AS_SA = case_when(
36   AS_SA == '[1000-10000>' ~ 'a.[1000-10000>',
37   AS_SA == '[10000-12000>' ~ 'b.[10000-12000>',
38   AS_SA == '[12000-15000>' ~ 'c.[12000-15000>',
39   AS_SA == '[15000-18000>' ~ 'd.[15000-18000>',
40   AS_SA == '[18000-24000>' ~ 'e.[18000-24000>',
41   AS_SA == '[24000-30000>' ~ 'f.[24000-30000>',
42   AS_SA == '[30000-36000>' ~ 'g.[30000-36000>',
43   AS_SA == '[36000-99999>' ~ 'h.[36000-99999]>'
44 ))
45 # Convertimos a factor las variables
46 data.Año <- as.factor(data.Año)
47 data.AS_Tipo_Poliza <- as.factor(data.AS_Tipo_Poliza)
48 data.AS_Antiguedad <- as.factor(data.AS_Antiguedad)
49 data.AS_Canal_Venta <- as.factor(data.AS_Canal_Venta)
50 data.AS_Clase_Veh <- as.factor(data.AS_Clase_Veh)
51 data.AS_Tipo_Veh <- as.factor(data.AS_Tipo_Veh)
52 data.AS_Uso_Veh <- as.factor(data.AS_Uso_Veh)
53 data.AS_Marca_Veh <- as.factor(data.AS_Marca_Veh)
54 data.AS_Region <- as.factor(data.AS_Region)
55 data.AS_Zona <- as.factor(data.AS_Zona)
56 data.AS_Genero <- as.factor(data.AS_Genero)
57 data.AS_Tipo_Persona <- as.factor(data.AS_Tipo_Persona)
58 data.AS_Edad <- as.factor(data.AS_Edad)
59 data.AS_SA <- as.factor(data.AS_SA)
60 # Asignamos el intercepto para cada variable
61 data <- data %>% mutate(
62   Año = fct_relevel(Año, "2022", after = 0),
63   AS_Tipo_Poliza = fct_relevel(AS_Tipo_Poliza, "Individual", after = 0),
64   AS_Antiguedad = fct_relevel(AS_Antiguedad, "b. 1-2", after = 0),
65   AS_Canal_Venta = fct_relevel(AS_Canal_Venta, "Corredores", after = 0),
66   AS_Clase_Veh = fct_relevel(AS_Clase_Veh, "Livialos", after = 0),
67   AS_Tipo_Veh = fct_relevel(AS_Tipo_Veh, "Automovil", after = 0),
68   AS_Uso_Veh = fct_relevel(AS_Uso_Veh, "Particular", after = 0),
69   AS_Marca_Veh = fct_relevel(AS_Marca_Veh, "L2", after = 0),
70   AS_Region = fct_relevel(AS_Region, "Lima", after = 0),
71   AS_Zona = fct_relevel(AS_Zona, "A3", after = 0),
72   AS_Genero = fct_relevel(AS_Genero, "Femenino", after = 0),
73   AS_Tipo_Persona = fct_relevel(AS_Tipo_Persona, "Empresa", after = 0),
74   AS_Edad = fct_relevel(AS_Edad, "38-42", after = 0),
75   AS_SA = fct_relevel(AS_SA, "a.[1000-10000>", after = 0)
76 )
77 # Verificación de frecuencia por factor
78 data %>% group_by(Año) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)

```

```

79 data %>% group_by(AS_Tipo_Poliza) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E
80   ,5)*100)
81 data %>% group_by(AS_Antiguedad) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)
82   *100)
83 data %>% group_by(AS_Canal_Venta) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E
84   ,5)*100)
85 data %>% group_by(AS_Clase_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
86   100)
87 data %>% group_by(AS_Tipo_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
88   100)
89 data %>% group_by(AS_Uso_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
90   100)
91 data %>% group_by(AS_Marca_Veh) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
92   100)
93 data %>% group_by(AS_Region) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
94   100)
95 data %>% group_by(AS_Zona) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
96 data %>% group_by(AS_Genero) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*
97   100)
98 data %>% group_by(AS_Tipo_Persona) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E
99   ,5)*100)
100 data %>% group_by(AS_Edad) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
101 data %>% group_by(AS_SA) %>% summarise(E=sum(Expuesto), C=sum(Frecuencia)) %>% mutate(Freq=round(C/E,5)*100)
102 # Agrupaciones motivadas por modelos previos
103 data <- data %>%
104   mutate(
105     AS_Antiguedad=fct_collapse(AS_Antiguedad,"c. 3+"=c("c. 3","d. 4-5","e. 6","f. 7","g. 8+")
106   ),
107     AS_Canal_Venta=fct_collapse(AS_Canal_Venta,"AgentesExc_BancaDigital"=c("Agentes Exclusivos","Banca y
108       Digital"))
109   ),
110     AS_Edad=fct_collapse(AS_Edad,"43+"=c("43-47","48-52","53-57","58-67","68-72","73-82","83+")
111   ),
112     AS_Genero=fct_collapse(AS_Genero,"Masculino"=c("Indeterminado","Masculino"))
113   ),
114     AS_Marca_Veh=fct_collapse(AS_Marca_Veh,"L3_L4"=c("L3","L4"), "M1_M2_M3"=c("M1","M2","M3"), "P3_P4"=c("P3",
115       "P4"), "P2_P5"=c("P2","P5"))
116   ),
117     AS_SA=fct_collapse(
118       AS_SA,"b.[10000-18000>"=c("b.[10000-12000>","c.[12000-15000>","d.[15000-18000>"),
119       "e.[18000-24000>","f.[24000-30000>","g.[30000-36000>","h.[36000-99999]")
120   ),
121     AS_Uso_Veh=fct_collapse(
122       AS_Uso_Veh,"Alquiler_Escolar"=c("Alquiler","Escolar"), "Taxi_Amb_Tur_Com"=c("Taxi","Ambulancia","Turismo"
123       ,"Comercial"), "TransPer_Carga_TransInt_TransUrb_Instr"=c("Transporte Personal","Carga","Transporte
124       Interprovincial","Transporte Urbano","Instrucción"))
125   ),
126     AS_Zona=fct_collapse(AS_Zona,"A1_A2"=c("A1","A2"), "A4_A5"=c("A4","A5"), "D1_D4_D5"=c("D1","D4","D5"),
127       "D2_D3"="c("D2","D3"))
128   )
129 #####
130 # Agrupamos la data para modelar el GLM
131 #data_model <- data %>% mutate(id=as.numeric('Número Póliza')*100+as.numeric('Número Riesgo')) %>% group_by(id
132   ,Año,PP_Tipo_Poliza,PP_Antiguedad,PP_Canal_Venta,PP_Uso_Veh,PP_Marca_Veh,PP_Zona,PP_Genero,PP_Edad,PP_SA)
133   %>% summarise(Expuesto=sum(Expuesto),Prima_Devengada=sum(Prima_Devengada),Cantidad=sum(Frecuencia),
134     Incurrido=sum(Incurrido))
135 data_model <- data %>% group_by(Año,
136   #PT_Tipo_Poliza,
137   AS_Antiguedad,
138   AS_Canal_Venta,
139   AS_Uso_Veh,
140   AS_Marca_Veh,
141   AS_Zona,
142   AS_SA
143 )%>% summarise(Expuesto=sum(Expuesto),Prima_Devengada=sum(Prima_Devengada),Cantidad=sum(Frecuencia),Incurrido=
144   sum(Incurrido))
145 # Modelo GLM Asistencias
146 glm.Frec.AS <- glm(
147   formula = Cantidad ~
148     Año +
149     #AS_Tipo_Poliza +
150     AS_Antiguedad +
151     AS_Canal_Venta +
152     #AS_Clase_Veh + (Se correlaciona con la MarcaVeh)
153     #AS_Tipo_Veh + (Se correlaciona con la MarcaVeh)

```

```

135 AS_Uso_Veh +
136 AS_Marca_Veh +
137 #AS_Region + (Se correlaciona con la zona)
138 AS_Zona +
139 #AS_Genero +
140 #AS_Tipo_Persona + (Se correlaciona con la Edad)
141 #AS_Edad +
142 AS_SA +
143 offset(log(Expuesto)),
144 family = poisson(link = "log"),
145 data = data_model
146 )
147 # Desenmascarar posibles multicolinealidades entre las diferentes variables
148 temp <- alias(glm.Frec.AS)
149 temp2 <- as.data.frame(temp.Complete) # Si es vacio, no hay multicolinealidades
150 rm(temp,temp2)
151 # No se realizarán agrupaciones en los niveles de algunas variables por tener un buen modelo
152 # Incluimos las predicciones en la data
153 data <- data %>% mutate(Frecuencia.PRED = glm.Frec.AS.fitted.values)
154 data %>% summarise(sum(Frecuencia), sum(Frecuencia.PRED))
155 # Salvar un data frame con los coeficientes del modelo
156 coefs.Frecuencia.AS <- summary(glm.Frec.AS).coefficients %>%
157   as_tibble() %>%
158   mutate(var_level = glm.Frec.AS.coefficients %>% names()) %>%
159   select(var_level, everything())
160 # Adicionar manualmente a la tabla los niveles atribuidos al intercept
161 # (y arredondar los valores)
162 temp <- tibble(
163   var_level = c("Año2022","AS_Antiguedadb_ 1-2","AS_Canal_VentaCorredores","AS_Uso_VehParticular","AS_Marca_VehL2","AS_ZonaA3","AS_SAA.[1000-10000]>",
164   ),
165   Estimate = 0
166 )
167 f.round <- function(x, .digits = 4) round(x, digits = .digits)
168 coefs.Frecuencia.AS <- coefs.Frecuencia.AS %>%
169   bind_rows(temp) %>%
170   arrange(var_level) %>%
171   mutate(`exp(Estimate)` = exp(Estimate)) %>%
172   mutate_if(is.double, f.round)
173 # Esta función considera los datos de entrenamiento del GLM, agrega la exposición por variable y
174 # la junta a la tabla de coeficientes.
175 incluir.exposicion <- function(.data, .tbl.coefs){
176   names.vars <- names(.data %>% select(-Expuesto))
177   n <- length(names.vars)
178   list.temp <- vector(mode = "list", length = n)
179   for(i in seq_along(names.vars)){
180     varname <- sym(names.vars[[i]])
181     list.temp[[i]] <- .data %>%
182       group_by(!!varname) %>%
183       summarize(Exposición = f.round(sum(Expuesto), .digits = 0)) %>%
184       ungroup() %>%
185       mutate(var_level = str_c(names.vars[[i]], !!varname)) %>%
186       select(var_level, Exposición)
187   }
188   exp.by.var <- bind_rows(list.temp)
189   return(.tbl.coefs %>% left_join(exp.by.var))
190 }
191 coefs.Frecuencia.AS <- data %>%
192   select(-Frecuencia, -Frecuencia.PRED) %>%
193   incluir.exposicion(coefs.Frecuencia.AS)
194 coefs.Frecuencia.AS %>% DT:::datatable()
195 # Guardamos los coeficientes en excel
196 write.csv(coefs.Frecuencia.AS,"C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/01_Frecuencia/04_Coeficientes/coefs.Frecuencia.AS.csv",fileEncoding ="Latin1")
197 save(glm.Frec.AS, file = "C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/glm.Frec.AS.RData")
198 save(data, file="C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/Data_AS.RData")
199 save(data_model, file="C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/Data_AS_model.RData")

```

Código R: Modelo Severidad - Asistencias

```

1 #####
2 ##### MODELO GLM – SEVERIDAD

```

```

3 ######
4 # PAQUETES Y LIBRERIAS
5 #options(repos = c(cran="http://cran.rstudio.com"))
6 #install.packages("RODBC")
7 #install.packages("skimr")
8 library(RODBC)
9 library(tidyverse) # Incluye dplyr y tidyverse para manejar los datos y ggplot para representarlos en gráficos
10 library(plotly) # Para graficos interactivos
11 library(lubridate) # manejar datos y horas
12 library(forcats)
13 library(rchartcolor)
14 library(readr) # para leer y importar archivos
15 library(skimr)
16 db2<-"/C:/Users/josephgarcia/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/02_CostoMedio/03_Data/BD_GLM_CM.acedb"
17 con2 <- odbcConnectAccess2007(db2)
18 #data_CM <- sqlQuery(con2,"select * from AS_Datos_Severidad")
19 #str(data_CM)
20 # Reducimos la data sin el año 2015 y 2020, de acuerdo a los análisis univariados realizados
21 #data_CM <- data_CM %>% filter(Año!=2015 & Año!=2020)
22 #data_CM <- data_CM %>% filter(AS_Clase_Veh!="Desconocido")
23 #data_CM <- data_CM %>% filter(Incurrido>0)
24 #data_CM.‘Número Póliza’<-format(data_CM.‘Número Póliza’,scientific = FALSE)
25 #data_CM <- data_CM %>% filter(Cantidad<100)
26 data_CM <- data_model %>% filter(Incurrido>0) # omitir todas las acciones y saltar al modelo GLM
27 # Cambiamos el nombre de algunos niveles de algunas variables
28 data_CM <- data_CM %>% mutate(AS_Antiguedad = case_when(
29   AS_Antiguedad == '0' ~ 'a. 0', AS_Antiguedad == '1-2' ~ 'b. 1-2', AS_Antiguedad == '3' ~ 'c. 3',
30   AS_Antiguedad == '4-5' ~ 'd. 4-5', AS_Antiguedad == '6' ~ 'e. 6', AS_Antiguedad == '7' ~ 'f. 7', AS_Antiguedad == '8+' ~ 'g. 8+'
31 ))
32 data_CM <- data_CM %>% mutate(AS_SA = case_when(
33   AS_SA == '[1000-10000>' ~ 'a.[1000-10000>', AS_SA == '[10000-12000>' ~ 'b.[10000-12000>', AS_SA == '[12000-15000>' ~ 'c.[12000-15000>', AS_SA == '[15000-18000>' ~ 'd.[15000-18000>', AS_SA == '[18000-24000>' ~ 'e.[18000-24000>', AS_SA == '[24000-30000>' ~ 'f.[24000-30000>', AS_SA == '[30000-36000>' ~ 'g.[30000-36000>', AS_SA == '[36000-99999>' ~ 'h.[36000-99999]'
34 ))
35 # Convertimos a factor las variables
36 data_CM.Año <- as.factor(data_CM.Año)
37 data_CM.AS.Tipo_Poliza <- as.factor(data_CM.AS.Tipo_Poliza)
38 data_CM.AS.Antiguedad <- as.factor(data_CM.AS.Antiguedad)
39 data_CM.AS.Canal_Venta <- as.factor(data_CM.AS.Canal_Venta)
40 data_CM.AS.Clase_Veh <- as.factor(data_CM.AS.Clase_Veh)
41 data_CM.AS.Tipo_Veh <- as.factor(data_CM.AS.Tipo_Veh)
42 data_CM.AS.Uso_Veh <- as.factor(data_CM.AS.Uso_Veh)
43 data_CM.AS.Mарка_Veh <- as.factor(data_CM.AS.Mарка_Veh)
44 data_CM.AS.Region <- as.factor(data_CM.AS.Region)
45 data_CM.AS.Zona <- as.factor(data_CM.AS.Zona)
46 data_CM.AS.Genero <- as.factor(data_CM.AS.Genero)
47 data_CM.AS.Tipo_Persona <- as.factor(data_CM.AS.Tipo_Persona)
48 data_CM.AS.Edad <- as.factor(data_CM.AS.Edad)
49 data_CM.AS.SA <- as.factor(data_CM.AS.SA)
50 # Asignamos el intercepto para cada variable
51 data_CM <- data_CM %>% mutate(Año = fct_relevel(Año, "2022", after = 0),
52   AS_Tipo_Poliza = fct_relevel(AS_Tipo_Poliza, "Individual", after = 0),
53   AS_Antiguedad = fct_relevel(AS_Antiguedad, "b. 1-2", after = 0),
54   AS_Canal_Venta = fct_relevel(AS_Canal_Venta, "Corredores", after = 0),
55   AS_Clase_Veh = fct_relevel(AS_Clase_Veh, "Livialos", after = 0),
56   AS_Tipo_Veh = fct_relevel(AS_Tipo_Veh, "Automovil", after = 0),
57   AS_Uso_Veh = fct_relevel(AS_Uso_Veh, "Particular", after = 0),
58   AS_Mарка_Veh = fct_relevel(AS_Mарка_Veh, "L2", after = 0),
59   AS_Region = fct_relevel(AS_Region, "Lima", after = 0),
60   AS_Zona = fct_relevel(AS_Zona, "A3", after = 0),
61   AS_Genero = fct_relevel(AS_Genero, "Femenino", after = 0),
62   AS_Tipo_Persona = fct_relevel(AS_Tipo_Persona, "Empresa", after = 0),
63   AS_Edad = fct_relevel(AS_Edad, "38-42", after = 0
64 ), AS_SA = fct_relevel(AS_SA, "a.[1000-10000]", after = 0)
65 )
66 # Verificación de frecuencia por factor
67 data_CM %>% group_by(Año) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
68 data_CM %>% group_by(AS_Tipo_Poliza) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
69 data_CM %>% group_by(AS_Antiguedad) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
70 data_CM %>% group_by(AS_Canal_Venta) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
71 data_CM %>% group_by(AS_Clase_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
72 data_CM %>% group_by(AS_Tipo_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
73 data_CM %>% group_by(AS_Uso_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))

```

```

74 data_CM %>% group_by(AS_Marca_Veh) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
75 data_CM %>% group_by(AS_Region) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
76 data_CM %>% group_by(AS_Zona) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
77 data_CM %>% group_by(AS_Genero) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
78 data_CM %>% group_by(AS_Tipo_Persona) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
79 data_CM %>% group_by(AS_Edad) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
80 data_CM %>% group_by(AS_SA) %>% summarise(M=sum(Incurrido), C=n()) %>% mutate(CM=round(M/C,5))
81 # Agrupaciones motivadas por modelos previos
82 data_CM <- data_CM %>%
83   mutate(
84     AS_Antiguedad=fct_collapse(AS_Antiguedad,"c. 3+"=c("c. 3","d. 4-5","e. 6","f. 7","g. 8+")),
85     AS_Canal_Venta=fct_collapse(AS_Canal_Venta,"AgentesExc_BancaDigital"=c("Agentes Exclusivos","Banca y
86       Digital")),
87     AS_Edad=fct_collapse(AS_Edad,"43+"=c("43-47","48-52","53-57","58-67","68-72","73-82","83+")),
88     AS_Genero=fct_collapse(AS_Genero,"Masculino"=c("Indeterminado","Masculino")),
89     AS_Marca_Veh=fct_collapse(AS_Marca_Veh,"L3_L4"=c("L3","L4"), "M1_M2_M3"=c("M1","M2","M3"), "P3_P4"=c("P3",
90       "P4"), "P2_P5"=c("P2","P5")),
91     AS_SA=fct_collapse(AS_SA,"b.[10000-18000>"=c("b.[10000-12000>","c.[12000-15000>","d.[15000-18000>"),
92       ".[18000-99999]"=c("e.[18000-24000>","f.[24000-30000>","g.[30000-36000>","h.[36000-999999"])),
93     AS_Uso_Veh=fct_collapse(AS_Uso_Veh,"Alquiler_Escolar"=c("Alquiler","Escolar"), "Taxi_Amb_Tur_Com"=c("Taxi",
94       "Ambulancia","Turismo","Comercial"), "TransPer_Carga_TransInt_TransUrb_Instr"=c("TransportePersonal",
95       "Carga","Transporte Interprovincial","Transporte Urbano")),
96     AS_Zona=fct_collapse(AS_Zona,"A1_A2"=c("A1","A2"), "A4_A5"=c("A4","A5"), "D1_D4_D5"=c("D1","D4","D5"), "D2_D3"
97       "=c("D2","D3"))
98   )
99 # Creamos la variable dependiente
100 data_CM <- data_CM %>% mutate(CM=Incurrido/Cantidad)
101 # Modelo GLM Pérdida Parcial
102 glm.Sev.AS <- glm(
103   formula = CM ~
104   Año +
105   #AS_Tipo_Poliza +
106   AS_Antiguedad +
107   AS_Canal_Venta +
108   #AS_Clase_Veh + (Se correlaciona con la MarcaVeh)
109   #AS_Tipo_Veh + (Se correlaciona con la MarcaVeh)
110   AS_Uso_Veh +
111   AS_Marca_Veh +
112   #AS_Region + (Se correlaciona con la zona)
113   AS_Zona +
114   #AS_Genero +
115   #AS_Tipo_Persona + (Se correlaciona con la Edad)
116   #AS_Edad,
117   AS_SA,
118   #offset(log(Expuesto)), Se coloca este offset en caso que la variable dependiente hubiera sido solo el
119   # Incurrido
120   family = Gamma(link = "log"),
121   data = data_CM
122 )
123 # Desenmascarar posibles multicolinearidades entre las diferentes variables
124 temp <- alias(glm.Sev.AS)
125 temp2 <- as.data.frame(temp.Complete) # Si es vacio, no hay multicolinearidades
126 rm(temp,temp2)
127 # No se realizarán agrupaciones en los niveles de algunas variables por tener un buen modelo
128 # Incluimos las predicciones en la data
129 data_CM <- data_CM %>% mutate(CM.PRED = glm.Sev.AS.fitted.values)
130 data_CM %>% summarise(sum(Incurrido), sum(CM.PRED*Cantidad))
131 # Salvar un data frame con los coeficientes del modelo
132 coefs.Severidad.AS <- summary(glm.Sev.AS).coefficients %>%
133   as_tibble() %>%
134   mutate(var_level = glm.Sev.AS.coefficients %>% names()) %>%
135   select(var_level, everything())
136 # Adicionar manualmente a la tabla los niveles atribuidos al intercept
137 # (y arredondar los valores)
138 temp <- tibble(
139   var_level = c("Año2022","AS_Antiguedadb. 1-2","AS_Canal_VentaCorredores","AS_Uso_VehParticular","AS_Marca_
140   VehL2","AS_ZonaA3","AS_SAa.[1000-10000>"),
141   Estimate = 0
142 )
143 f.round <- function(x, .digits = 4) round(x, digits = .digits)
144 coefs.Severidad.AS <- coefs.Severidad.AS %>%
145   bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate)` = exp(Estimate)) %>% mutate_if(is.double,
146   f.round)
147 # Esta función considera los datos de entrenamiento del GLM, agrega la exposición por variable y
148 # la junta a la tabla de coeficientes.

```

```

141 incluir.Cantidad <- function(.data, .tbl.coefs){
142   names.vars <- names(.data %>% select(-Cantidad))
143   n <- length(names.vars)
144   list.temp <- vector(mode = "list", length = n)
145   for(i in seq_along(names.vars)){
146     varname <- sym(names.vars[[i]])
147     list.temp[[i]] <- .data %>%
148       group_by(!varname) %>%
149       summarize(Incurrido = f.round(sum(Cantidad), .digits = 0)) %>%
150       ungroup() %>%
151       mutate(var_level = str_c(names.vars[[i]], !varname)) %>%
152       select(var_level, Incurrido)
153   }
154   exp.by.var <- bind_rows(list.temp)
155   return(.tbl.coefs %>% left_join(exp.by.var))
156 }
157 coefs.Severidad.AS <- data_CM %>%
158   select(-CM.PRED) %>%
159   incluir.Cantidad(coefs.Severidad.AS)
160 coefs.Severidad.AS %>% DT::datatable()
161 # Guardamos los coeficientes en excel
162 write.csv(coefs.Severidad.AS,"C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/02_CostoMedio/04_Coeficientes/coefs.Severidad.AS.csv",fileEncoding = "Latin1")
163 save(glm.Sev.AS, file = "C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/glm.Sev.AS.RData")
164 save(data_CM, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/Data_AS_model_CM.RData")

```

Código R: Índice de GINI

```

1 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Parcial/
      glm.Frec.PP.RData")
2 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Total/
      glm.Frec.PT.RData")
3 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Responsabilidad_Civil/
      glm.Frec.RC.RData")
4 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistancia/glm.
      Frec.AS.RData")
5
6 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Parcial/
      Data_PP.RData")
7 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Total/
      Data_PT.RData")
8 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Responsabilidad_Civil/
      Data_RC.RData")
9 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistancia/Data_AS.RData")
10
11 library(REAT)
12 library(ineq)
13
14 # PERDIDA PARCIAL
15 data <- data %>% mutate(FREC_PRED_PP=predict(glm.Frec.PP,newdata =data ,type = "response"))
16 data_PP<-cbind(data.Expuesto,data.FREC_PRED_PP)
17 data_PP <- as.data.frame(data_PP)
18 data <- data %>% mutate(FREC_PRED_PT=predict(glm.Frec.PT,newdata =data ,type = "response"))
19 # PERDIDA TOTAL
20 data_PT<-cbind(data.Expuesto,data.FREC_PRED_PT)
21 data_PT <- as.data.frame(data_PT)
22 # RESPONSABILIDAD CIVIL
23 data <- data %>% mutate(FREC_PRED_RC=predict(glm.Frec.RC,newdata =data ,type = "response"))
24 data_RC<-cbind(data.Expuesto,data.FREC_PRED_RC)
25 data_RC <- as.data.frame(data_RC)
26 # ASISTENCIA
27 data <- data %>% mutate(FREC_PRED_AS=predict(glm.Frec.AS,newdata =data ,type = "response"))
28 data_AS<-cbind(data.Expuesto,data.FREC_PRED_AS)
29 data_AS <- as.data.frame(data_AS)
30 # GRAFICOS
31 par(mfrow=c(2,2))
32 lorenz(data_PP.V2 ,weighting = data_PP.VI,leg = TRUE,lcx = "Porcentaje de Expuestos",lcy="Porcentaje de Pérdidas",lctitle="Índice de Gini - Pérdida Parcial",lsize=2)
33 lorenz(data_PT.V2 ,weighting = data_PT.VI,leg = TRUE,lcx = "Porcentaje de Expuestos",lcy="Porcentaje de Pérdidas",lctitle="Índice de Gini - Pérdida Total",lsize=2)
34 lorenz(data_RC.V2 ,weighting = data_RC.VI,leg = TRUE,lcx = "Porcentaje de Expuestos",lcy="Porcentaje de Pérdidas",lctitle="Índice de Gini - Responsabilidad Civil",lsize=2)

```

```
35 | lorenz(data_AS.V2 ,weighting = data_AS.V1,lcg = TRUE,lcx = "Porcentaje de Expuestos",lcy="Porcentaje de Pérdidas",lctitle="Índice de Gini - Asistencias",lsize=2)
```

ANEXO C: INDICIO DE DEPENDENCIA Y MODELOS GLM CONJUNTOS

C.1. Indicios de Dependencia

```

1 data_PP <- data %>%
2   group_by(Año, PP_Tipo_Poliza, PP_Antiguedad, PP_Canal_Venta, PP_Uso_Veh, PP_Marca_Veh, PP_Zona, PP_Genero, PP_Edad,
3             PP_SA) %>%
4   summarise(Exp=sum(Expuesto), Cantidad=sum(Frecuencia), Incurrido=sum(Incurrido))
5
6 data_PP <- data_PP %>% mutate(
7   Frecuencia=Cantidad/Exp,
8   Severidad=Incurrido/Cantidad
9 )
10
11 data_PP <- data_PP %>% filter(Cantidad>0 & Severidad>=0)
12 write.csv(data_PP, "C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/10_Overleaf_Tesis/Cuadros/Indicio_"
13            DEP.csv", fileEncoding = "Latin1")
14 #
15 data_PT <- data %>%
16   group_by(Año, PT_Tipo_Poliza, PT_Uso_Veh, PT_Marca_Veh, PT_Zona, PT_Genero, PT_Edad) %>%
17   summarise(Exp=sum(Expuesto), Cantidad=sum(Frecuencia), Incurrido=sum(Incurrido))
18
19 data_PT <- data_PT %>% mutate(
20   Frecuencia=Cantidad/Exp,
21   Severidad=Incurrido/Cantidad
22 )
23 data_PT <- data_PT %>% filter(Cantidad>0 & Severidad>=0)
24 write.csv(data_PT, "C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/10_Overleaf_Tesis/Cuadros/Indicio_"
25            DEP_PT.csv", fileEncoding = "Latin1")
26 #
27 data_RC <- data %>%
28   group_by(Año, RC_Antiguedad, RC_Canal_Venta, RC_Uso_Veh, RC_Marca_Veh, RC_Zona, RC_Edad) %>%
29   summarise(Exp=sum(Expuesto), Cantidad=sum(Frecuencia), Incurrido=sum(Incurrido))
30
31 data_RC <- data_RC %>% mutate(
32   Frecuencia=Cantidad/Exp,
33   Severidad=Incurrido/Cantidad
34 )
35 data_RC <- data_RC %>% filter(Cantidad>0 & Severidad>=0)
36 write.csv(data_RC, "C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/10_Overleaf_Tesis/Cuadros/Indicio_"
37            DEP_RC.csv", fileEncoding = "Latin1")
38 #
39 data_AS <- data %>%
40   group_by(Año, AS_Antiguedad, AS_Canal_Venta, AS_Uso_Veh, AS_Marca_Veh, AS_Zona, AS_SA) %>%
41   summarise(Exp=sum(Expuesto), Cantidad=sum(Frecuencia), Incurrido=sum(Incurrido))
42
43 data_AS <- data_AS %>% mutate(
44   Frecuencia=Cantidad/Exp,
45   Severidad=Incurrido/Cantidad
46 )
47 data_AS <- data_AS %>% filter(Cantidad>0 & Severidad>=0)
48 write.csv(data_AS, "C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/10_Overleaf_Tesis/Cuadros/Indicio_"
49            DEP_AS.csv", fileEncoding = "Latin1")
50 #

```

C.2. Gráficos CHI-PLOT y K-PLOT

```

1
2 library(MASS)
3 library(evd)
4 library(asbio)
5 library(copula)
6 library(VineCopula)
7 library(base)
8 load("C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Parcial/"
9      glm.Frec.PP.RData")
10 load("C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Total/"
11      glm.Frec.PT.RData")
12 load("C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Responsabilidad_Civil/"
13      glm.Frec.RC.RData")
14 load("C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/glm."
15      Frec.AS.RData")

```

```

12 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Parcial/
13 Data_PP.RData")
14 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Total/
15 Data_PT.RData")
16 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Responsabilidad_
17 Civil/Data_RC.RData")
18 load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia/
19 Data_AS.RData")
20
21 FS_PP<- cbind(data_PP.Cantidad,data_PP.Severidad)
22 FS_PP<-pobs(FS_PP)
23 FS_PP <- as.data.frame(FS_PP)
24 rand_df_PP <- FS_PP[sample(nrow(FS_PP), size=1000), ]
25
26 FS_PT<- cbind(data_PT.Cantidad,data_PT.Severidad)
27 FS_PT<-pobs(FS_PT)
28 FS_PT <- as.data.frame(FS_PT)
29 rand_df_PT <- FS_PT[sample(nrow(FS_PT), size=1000), ]
30
31 FS_RC<- cbind(data_RC.Cantidad,data_RC.Severidad)
32 FS_RC<-pobs(FS_RC)
33 FS_RC <- as.data.frame(FS_RC)
34 rand_df_RC <- FS_RC[sample(nrow(FS_RC), size=1000), ]
35
36 FS_AS <- cbind(data_AS.Cantidad,data_AS.Severidad)
37 FS_AS <-pobs(FS_AS)
38 FS_AS <- as.data.frame(FS_AS)
39 rand_df_AS <- FS_AS[sample(nrow(FS_AS), size=1000), ]
40
41 par(mfrow=c(2,2))
42 BiCopChiPlot(rand_df_PP.V1,rand_df_PP.V2, ylim=c(-0.1,1), main="Pérdida Parcial", col=4)
43 BiCopChiPlot(rand_df_PT.V1,rand_df_PT.V2, ylim=c(-0.1,1), main="Pérdida Total", col=4)
44 BiCopChiPlot(rand_df_RC.V1,rand_df_RC.V2, ylim=c(-0.1,1), main="Responsabilidad Civil", col=4)
45 BiCopChiPlot(rand_df_AS.V1,rand_df_AS.V2, ylim=c(-0.1,1), main="Asistencias", col=4)
46
47 BiCopKPlot(rand_df_PP.V1,rand_df_PP.V2, lwd=2, col=4, main="Pérdida Parcial")
48 BiCopKPlot(rand_df_PT.V1,rand_df_PT.V2, lwd=2, col=4, main="Pérdida Total")
49 BiCopKPlot(rand_df_RC.V1,rand_df_RC.V2, lwd=2, col=4, main="Responsabilidad Civil")
50 BiCopKPlot(rand_df_AS.V1,rand_df_AS.V2, lwd=2, col=4, main="Asistencias")

```

C.3. Modelo Conjunto - Pérdida Parcial

```

1 ##########
2 ##########
3 #####
4 ##### REGRESION BASADO EN COPULAS #####
5 ##### PERDIDA PARCIAL #####
6 #####
7 ##########
8 ##########
9 # Cargamos las librerías necesarias
10 library(copula)
11 library(RODBC)
12 library(tidyverse)
13 library(MASS)
14 #library(GRM)
15 #library(devtools)
16 library(CopulaRegression)
17 library(VineCopula)
18 library(optimx)
19 #install_url('http://cran.r-project.org/src/contrib/Archive/CopulaRegression/CopulaRegression_0.1-5.tar.gz')
20 #####
21 # REGRESION GAUSSIANA
22 #####
23
24 # Set Directory
25 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_
26 Parcial")
27
28 # Carga los modelos GLM MARGINALES calculados en otros queries
29 load("glm.Frec_PP.RData")
30 load("glm.Sev_PP.RData")
31 load("Data_PP.RData")
32 load("Data_PP_model.RData")

```

```

32 load("Data_PP_model_CM.RData")
33
34 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
35 # SEVERIDAD
36 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
37 var_x <- model.matrix(glm.Sev.PP) #data_CM[,3:12]
38
39 # FRECUENCIA
40 y <- data_model.Cantidad
41 Expuestos <- data_model.Expuesto
42 var_y <- model.matrix(glm.Frec.PP) #data[,3:12]
43
44 # Estimacion de parámetros iniciales utilizando las distribuciones marginales
45 # SEVERIDAD
46 sd_alpha <- sqrt(diag(vcov(glm.Sev.PP)))
47 alpha_0 <- glm.Sev.PP.coefficients
48 delta_0 <- summary(glm.Sev.PP).dispersion
49 mu_0 <- exp(var_x%*%alpha_0)
50
51 # FRECUENCIA
52 beta_0 <- glm.Frec.PP.coefficients
53 sd_beta <- sqrt(diag(vcov(glm.Frec.PP)))
54 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
55
56 # REGRESION MEDIANTE COPULAS
57
58 # COPULA GAUSSIANA
59 family=1
60 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MLE
61 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-lambda_0)/(length(y)+1), family=family).par
62 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
63 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
64
65 # Estimación de las seudo-observaciones
66 u<-pgam(x,mu_0,delta_0)
67 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
68 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
69
70 # Creamos la función primera derivada de la Función Copula
71 Du_Gaus <- function(u,v,theta){
72   u[u<=0]=0.001
73   u[u>=1]=0.999
74   v[v<=0]=0.001
75   v[v>=1]=0.999
76   out <- pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
77   return(out)
78 }
79
80 # Función a maximizar
81 f_aux <- function(para){
82   D_u_0<-Du_Gaus(u,v,para)-Du_Gaus(u,vv,para)
83   D_u_0[D_u_0<=0]=1
84   D_u_0 <- log(D_u_0)
85   out<-(-sum(D_u_0))
86   return(out)
87 }
88
89 # Proceso de maximización
90 para_ini <- theta2z(theta_ini,family = family)
91 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
92 #theta_0 <- z2theta(para_ifm,family = family)
93 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
# para la regresión conjunta
94 tau <- BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
95
96 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
97
98 # Creamos nuestra función a maximizar
99 #####
f_aux_reg <- function(para){
100
101   p <- ncol(var_x)
102   q <- ncol(var_x)
103   alpha <- para[1:p]
104   beta <- para[(p+1):(p+q)]
```

```

105  # theta  <- z2theta(para[p+q+1],family)
106 theta  <- para[p+q+1]
107 delta  <- para[p+q+2]
108 lambda <- as.vector(exp(var_y%%beta)*data_model.Expuesto)
109 mu     <- as.vector(exp(var_y%%alpha))
110 #dummy  <- density_joint(x,y,mu,delta,theta,family=family,zt=FALSE)
111 u2 <- pgam(data_model.Incurrido,mu,delta)
112 v2<-ppois(y,lambda)
113 vv2<-ppois(y-1,lambda)
114
115 marginal.x <- dgam(data_model.Incurrido,mu,delta)
116 marginal.x[marginal.x>=1]=0
117 marginal.x[marginal.x<=0]=0
118 marginal.x[is.na(marginal.x)]=0
119 marginal.x[marginal.x==Inf]=0
120 marginal.x[marginal.x== -Inf]=0
121 par_der  <- Du_Gaus(u2,v2,theta)
122 par_der1 <- Du_Gaus(u2,vv2,theta)
123 dummy<- par_der-par_der1
124 dummy[y==0]=par_der[y==0]
125 out<-marginal.*dummy
126 out[out<=0]=1e-10
127 ll      <- -sum(log(out))
128 #if(negative==TRUE) ll <- (-ll)
129 return(ll)
130 }
131 #####
132 para_0<-c(alpha_0,beta_0,theta_0,delta_0) # Parámetros iniciales
133 a<-now()
134 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = FALSE) # Estimación Maxima verosmil de los
135           # parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
136 b<-now()
137 b-a
138 #load("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
139          Parcial/JointModel_Gaus.RData")
140 para_optim.value # Log_Likelihood
141 vector <- para_optim.par
142 p       <- ncol(var_y)
143 q       <- ncol(var_y)
144
145 # Parámetros optimizados
146 alpha_gau<-vector[1:p]
147 beta_gau<-vector[(p+1):(p+q)]
148 theta_gau=z2theta(vector[p+q+1],family) #-0.09506928
149 delta_gau<-exp(vector[p+q+2]) #4.271691
150 tau_gau<-BiCopPar2Tau(par=theta_gau,family=family)#-0.06061453
151 head(beta_0)
152 # Determinamos los errores estandar para el test de Wald
153 hessian_gau <- para_optim.hessian
154 Hinv        <- ginv(hessian_gau)
155 sd          <- sqrt(diag(Hinv))
156 sd.alpha_gau <-sd[1:p]
157 sd.beta_gau <- sd[(p+1):(p+q)]
158 sd.theta_gau <-sd[p+q+1]
159
160 # Ordenamos los coeficientes estimados en una tabla
161 Reg_Cop_Gausiana <- data.frame(Estimate_Freq=round(beta_gau,4),Std.error.freq=0,z_value_freq=0,"Prob_
162 freq"=0,Estimate_Sev=round(alpha_gau,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
163 var_level<-row.names(Reg_Cop_Gausiana)
164 Reg_Cop_Gausiana <- Reg_Cop_Gausiana %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_
165 level)
166
167 # Agregamos los interceptos
168 temp <- tibble(
169   var_level = c(
170     "Año2022",
171     "PP_Tipo_PolizaIndividual",
172     "PP_Antiguedad_c_02",
173     "PP_Canal_VentaCorredores",
174     "PP_Uso_Vehicular",
175     "PP_Marca_VehL3",
176     "PP_ZonaA2",
177     "PP_GeneroFemenino",
178     "PP_Edad38-42",
179     "PP_SAj.[13000-999999]")

```

```

177 ),
178 Estimate_Frec = 0
179 ,
180 Estimate_Sev = 0
181 )
182
183 # coeficientes del modelo de regresion con dependencia
184 Reg_Cop_Gausiana <- Reg_Cop_Gausiana %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)`=round(exp(Estimate_Frec),4), `exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
185 Reg_Cop_Gausiana <- Reg_Cop_Gausiana %>% relocate(`exp(Estimate_Frec)`, .after=Prob_frec)
186
187 # Guardamos el modelo conjunto y los coeficientes
188 write.csv(Reg_Cop_Gausiana, file="C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/Coef_Gaus.csv", fileEncoding = "Latin1")
189 save(para_optim, file="C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/JointModel_Gaus.RData")
190
191 # Simulacion de la Pérdida Total
192 # =====
193 # CON DEPENDENCIA
194 #
195 lambda_loss_dep <- exp(var_x%>%beta_gau)*data_CM.Expuesto
196 mu_loss_dep <- exp(var_x%>%alpha_gau)
197 delta_loss_dep <- delta_gau
198 # Función de densidad fx
199 f_loss <-
200   function(loss,mu,d,lambda){
201     n<-length(loss)
202     if (length(lambda)==1) lambda <- rep(lambda,n)
203     if (length(mu)==1) mu <- rep(mu,n)
204     out <- vector(length=n)
205
206     for(i in 1:n){
207       N <-1:20
208       v <-rpois(N,lambda[i])
209       vv <-rpois(N-1,lambda[i])
210       u <-pgam(loss[i]/N,mu[i],d)
211
212       Der_cop <-Du_Gaus(u,v,theta)-Du_Gaus(u,vv,theta)
213       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N
214       out[i] <-sum(dummy)
215     }
216
217     out[loss<=0]=0
218     return(out)
219   }
220
221 # Función de densidad acumulada Fx
222 F_loss <-
223   function(loss,mu,d,lambda){
224     out<-vector(length=length(loss))
225
226     for(i in 1:length(loss)){
227       floss <- function(s){
228         f_loss(s,mu[i],d,lambda[i],theta)
229       }
230       out[i] <- integrate(floss,0,loss[i]).value
231     }
232     return(out)
233   }
234
235 # Estimación del Total Loss (fuerte dedicacion computacional)
236
237 #k <-length(lambda_loss_dep)
238 k <- length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
239 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
240 L <- vector(length=k)
241 S <- vector(length=m)
242
243 for(j in 1:m){
244   r_uni<-runif(k)
245   for(i in 1:k){
246     f_root <-function(s){
247       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_gau)-r_uni[i]
248     }
249

```

```

250
251     tryCatch(
252       error = function(nd) mu_loss_dep[i]*lambda_loss_dep[i],
253       loss<-uniroot(f_root, lower = 0, upper = 500000)
254     )
255
256     #print(loss.root)
257     L[i]<-loss.root
258     perc<-paste(i/k*100,"%")
259     print(perc)
260
261     S[j]<-sum(L)
262     perc<-paste(j/m*100,"%")
263     print(perc)
264     #print(S[j])
265   }
266
267   # CON INDEPENDENCIA
268   # -----
269   lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN
270   # MODELO DE RIESGO COLECTIVO
271   mu_loss <-predict(glm.Sev.PP, newdata=data_CM, type='response')
272   delta_loss <-delta_0
273
274   M<-seq(1:10000)
275   for(i in 1:10000){
276     muestra_cant<-rpois(length(lambda_loss),lambda_loss)
277     S<-vector(length=length(lambda_loss))
278     for(j in 1:length(muestra_cant)){
279       if(muestra_cant[j]==0){
280         S[j]=0
281       } else {
282         S[j]=sum(rgam(n=muestra_cant[j],mu_loss[j],delta_loss))
283       }
284
285       M[i] <- sum(S)
286       porc<-paste((i/10000)*100,"%")
287       print(porc)
288       #print(M[i])
289     }
290
291     hist(M)
292   # Simulacion de la Frecuencia con Independencia
293   #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
294   # Frecuencia
295   N<-seq(1:10000)
296   for(i in 1:10000){
297     N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response")$.sim_1)/sum(data_model.Expuesto)
298     porc<-paste((i/10000)*100,"%")
299     print(porc)
300   }
301   hist(N)
302 #####
303   # Limpiamos el ambiente de variables
304   rm(list = ls())
305   gc()
306
307 #####
308   # REGRESION CLAYTON
309 #####
310
311   # Set Directory
312   setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_
313   Parcial")
314
315   # Carga los modelos GLM MARGINALES calculados en otros querys
316   load("glm.Frec.PP.RData")
317   load("glm.Sev.PP.RData")
318   load("Data_PP.RData")
319   load("Data_PP_model.RData")
320   load("Data_PP_model_CM.RData")
321
322   # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
323   # SEVERIDAD
324   x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)

```

```

324 var_x <- model.matrix(glm.Sev.PP) #data_CM[,3:12]
325 # FRECUENCIA
326 y <- data_model.Cantidad
327 Expuestos <- data_model.Expuesto
328 var_y <- model.matrix(glm.Frec.PP) #data[,3:12]
329
330 # Estimación de parámetros iniciales utilizando las distribuciones marginales
331 # SEVERIDAD
332 sd.alpha <- sqrt(diag(vcov(glm.Sev.PP)))
333 alpha_0 <- glm.Sev.PP.coefficients
334 delta_0 <- summary(glm.Sev.PP).dispersion
335 mu_0 <- exp(var_x%*%alpha_0)
336
337 # FRECUENCIA
338 beta_0 <- glm.Frec.PP.coefficients
339 sd.beta <- sqrt(diag(vcov(glm.Frec.PP)))
340 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
341
342 # REGRESION MEDIANTE COPULAS
343
344 # COPULA CLAYTON
345 family=3
346 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
347 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
lambda_0)/(length(y)+1), family=family).par
348 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
349 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
350
351 # Estimación de las seudo-observaciones
352 u<-pgam(x,mu_0,delta_0)
353 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
354 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
355
356 # Creamos la función primera derivada de la Función Copula
357 Du_Clay <- function(u,v,theta){
358   u[u<=0]=0.0001
359   u[u>1]=1
360   v[v<=0]=0.0001
361   v[v>1]=1
362   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
363   out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta)-1)
364   return(out)
365 }
366
367 # Función a maximizar
368 f_aux <- function(para){
369   D_u_0<-Du_Clay(u,v,para)-Du_Clay(u,vv,para)
370   D_u_0[D_u_0<=0]=1
371   D_u_0 <- log(D_u_0)
372   out<-(-sum(D_u_0))
373   return(out)
374 }
375
376 # Proceso de maximización
377 #para_ini <- theta2z(theta_ini,family = family)
378 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
379 #theta_0 <- z2theta(para_ifm,family = family)
380 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
# para la regresión conjunta
381 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
382
383 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
384
385 # Creamos nuestra función a maximizar
386 ##########
387 f_aux_reg <- function(para){
388   p <- ncol(var_x)
389   q <- ncol(var_x)
390   alpha <- para[1:p]
391   beta <- para[(p+1):(p+q)]
392   # theta <- z2theta(para[p+q+1],family = family)
393   theta <- z2theta(para[p+q+1],family = family)
394   delta <- para[p+q+2]
395   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
396   mu <- as.vector(exp(var_y%*%alpha))

```

```

397      #dummy  <- density_joint(x,y,mu,delta ,lambda ,theta ,family=family , zt=FALSE)
398      u2 <- pgam(data_model.Incurrido ,mu, delta )
399      v2<-ppois(y, lambda )
400      vv2<-ppois(y-1,lambda )
401
402      marginal.x <- dgam(data_model.Incurrido ,mu, delta )
403      marginal.x[marginal.x>=1]=0
404      marginal.x[marginal.x<=0]=0
405      marginal.x[is.na(marginal.x)]=0
406      marginal.x[marginal.x==Inf]=0
407      marginal.x[marginal.x== -Inf]=0
408      par_der  <- Du_Clay(u2,v2,theta )
409      par_der1 <- Du_Clay(u2,vv2,theta )
410      dummy<- par_der-par_der1
411      dummy[y==0]=par_der[y==0]
412      out<-marginal.x*dummy
413      out[out<=0]=1e-10
414      ll     <- -sum(log(out))
415      #if(negative==TRUE) ll <- (-ll)
416      return(ll)
417  }
418 #####
419 para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
420 a<-now()
421 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = FALSE) # Estimación Maxima verosmil de los
422      parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
423 b<-now()
424 b-a
425 #load("C:/Users/josephgarcial/OneDrive – KPMG/Tesis _Actuarial /03 _Modelos_GLM/06_GLM_Conjuntos /Perdida_
426      Parcial/JointModel_Gaus.RData")
427 para_optim.value # Log_Likelihood
428 vector <- para_optim.par
429 p       <- ncol(var_y)
430 q       <- ncol(var_y)
431 Log_Likelihood<-para_optim.value
432 # Parámetros optimizados
433 alpha_clay<-vector[1:p]
434 beta_clay<-vector[(p+1):(p+q)]
435 theta_clay<-z2theta(vector[p+q+1],family) #-0.09506928
436 delta_clay<-exp(vector[p+q+2]) #4.271691
437 tau_clay<-BiCopPar2Tau(par=theta_clay ,family=family) #-0.06061453
438 head(beta_clay)
439 # Determinamos los errores estandar para el test de Wald
440 hessian_gau <- para_optim.hessian
441 Hinvt     <- ginv(hessian_gau)
442 sd        <- sqrt(diag(Hinvt))
443 sd.alpha_gau <-sd[1:p]
444 sd.beta_gau <- sd[(p+1):(p+q)]
445 sd.theta_gau <-sd[p+q+1]
446
447 # Ordenamos los coeficientes estimados en una tabla
448 Reg_Cop_Gausiana <- data.frame(Estimate_Frec=round(beta_gau,4),Std.error.freq=round(sd.beta_gau,4),z_
449      value_frec=round(beta_gau/sd.beta_gau,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_gau/sd.beta_gau))),4),
450      Estimate_Sev=round(alpha_gau,4),Std.error.sev=round(sd.alpha_gau,4),z_value_sev=round(alpha_gau/sd.alpha_
451      gau,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_gau/sd.alpha_gau))),4))
452 Reg_Cop_Clayton <- data.frame(Estimate_Frec=round(beta_clay,4),Std.error.freq=0,z_value_frec=0,"Prob_frec"
453      =0,Estimate_Sev=round(alpha_clay,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
454 var_level<-row.names(Reg_Cop_Clayton)
455 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level
456      )
457
458 # Agregamos los interceptos
459 temp <- tibble(
460      var_level = c(
461          "Año2022",
462          "PP_Tipo_PolizaIndividual",
463          "PP_Antiguedad_c_02",
464          "PP_Canal_VentaCorredores",
465          "PP_Uso_VehParticular",
466          "PP_Marca_VehL3",
467          "PP_ZonaA2",
468          "PP_GeneroFemenino",
469          "PP_Edad38-42",
470          "PP_SAj.[13000-999999]"
471      ),
472      Estimate_Frec = 0

```

```

466      ,
467      Estimate_Sev = 0
468    )
469
470  # coeficientes del modelo de regresion con dependencia
471 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)` = round(exp(Estimate_Frec),4), `exp(Estimate_Sev)` = round(exp(Estimate_Sev),4)))
472 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% relocate(`exp(Estimate_Frec)`, .after=Prob_frec)
473
474 # Guardamos el modelo conjunto y los coeficientes
475 write.csv(Reg_Cop_Clayton, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/Coef_Clay.csv", fileEncoding = "Latin1")
476 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/JointModel_Clay.RData")
477
478 # Simulacion de la Pérdida Total
479 # =====
480
481 # CON DEPENDENCIA
482 # -----
483 lambda_loss_dep <- exp(var_x%*%beta_clay)*data_CM.Expuesto
484 mu_loss_dep <- exp(var_x%*%alpha_clay)
485 delta_loss_dep <- delta_clay
486 # Función de densidad fx
487
488 f_loss <-
489   function(loss ,mu,d, lam ,theta){
490     n<-length(loss)
491     if (length(lam)==1) lam <- rep(lam,n)
492     if (length(mu)==1) mu <- rep(mu,n)
493     out <- vector(length=n)
494
495     for(i in 1:n){
496       N <-1:20
497       v <-ppois(N, lam[i])
498       vv <-ppois(N-1, lam[i])
499       u <-pgam(loss[i]/N,mu[i],d)
500
501       Der_cop <-Du_Clay(u,v, theta)-Du_Clay(u,vv, theta)
502       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N
503       out[i] <-sum(dummy)
504     }
505
506     out[loss <=0]=0
507     return(out)
508   }
509
510 # Función de densidad acumulada Fx
511 F_loss <-
512   function(loss ,mu,d, lam ,theta){
513     out<-vector(length=length(loss))
514
515     for(i in 1:length(loss)){
516       floss <- function(s){
517         f_loss(s,mu[i],d, lam[i],theta)
518       }
519       out[i] <- integrate(floss ,0 ,loss[i]).value
520     }
521     return(out)
522   }
523
524 # Estimación del Total Loss (fuerte dedicacion computacional)
525
526 #k <-length(lambda_loss_dep)
527 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
528 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
529 L <- vector(length=k)
530 S <- vector(length=m)
531
532 for(j in 1:m){
533   r_uni<-runif(k)
534   for(i in 1:k){
535     f_root <-function(s){
536       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_clay)-r_uni[i]
537     }
538

```

```

539     tryCatch(
540       error = function(cond) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
541       loss<-uniroot(f_root, lower = 0, upper = 500000).root
542     )
543
544     #print(loss.root)
545     L[i]<-loss
546     perc<-paste(i/k*100,"%")
547     print(perc)
548   }
549   S[j]<-sum(L)
550   perc<-paste(j/m*100,"%")
551   print(perc)
552   #print(S[j])
553 }
554 S
555
556 # CON INDEPENDENCIA
557 # -----
558 lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO
559 mu_loss <-predict(glm.Sev.PP, newdata=data_CM, type='response')
delta_loss <-delta_0
560
561 M<-seq(1:10000)
562 for(i in 1:10000){
563   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
564   S<-vector(length=length(lambda_loss))
565   for(j in 1:length(muestra_cant)){
566     if(muestra_cant[j]==0){
567       S[j]=0
568     } else{
569       S[j]=sum(rgam(n=muestra_cant[j],mu_loss[j],delta_loss ))
570     }
571   }
572
573   M[i] <- sum(S)
574   porc<-paste((i/10000)*100,"%")
575   print(porc)
576   #print(M[i])
577 }
578
579 hist(M)
580 # Simulacion de la Frecuencia con Independencia
581 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
582 # Frecuencia
583 N<-seq(1:10000)
584 for(i in 1:10000){
585   N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
586   porc<-paste((i/10000)*100,"%")
587   print(porc)
588 }
589 hist(N)
590
591 #####
592 ##### Limpiamos el ambiente de variables
593 rm(list = ls())
594 gc()
595
596 #####
597 ##### REGRESION GUMBEL
598 ##### Set Directory
599 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Parcial")
600
601 # Carga los modelos GLM MARGINALES calculados en otros querys
602 load("glm.Frec.PP.RData")
603 load("glm.Sev.PP.RData")
604 load("Data_PP.RData")
605 load("Data_PP_model.RData")
606 load("Data_PP_model_CM.RData")
607
608 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal

```

```

613 # SEVERIDAD
614 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
615 var_x <- model.matrix(glm.Sev.PP) #data_CM[,3:12]
616 # FRECUENCIA
617 y <- data_model.Cantidad
618 Expuestos <- data_model.Expuesto
619 var_y <- model.matrix(glm.Frec.PP) #data[,3:12]
620
621 # Estimacion de parámetros iniciales utilizando las distribuciones marginales
622 # SEVERIDAD
623 sd_alpha <- sqrt(diag(vcov(glm.Sev.PP)))
624 alpha_0 <- glm.Sev.PP.coefficients
625 delta_0 <- summary(glm.Sev.PP).dispersion
626 mu_0 <- exp(var_x%*%alpha_0)
627
628 # FRECUENCIA
629 beta_0 <- glm.Frec.PP.coefficients
630 sd_beta <- sqrt(diag(vcov(glm.Frec.PP)))
631 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
632
633 # REGRESION MEDIANTE COPULAS
634
635 # COPULA GUMBEL
636 family=4
637 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
638 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
  lambda_0)/(length(y)+1), family=family).par
639 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
  data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
640 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
641
642 # Estimación de las seudo-observaciones
643 u<-pgam(x,mu_0,delta_0)
644 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
645 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
646
647 # Creamos la función primera derivada de la Función Copula
648 Du_Gum <- function(u,v,theta){
649   u[u<=0]=0.0001
650   u[u>1]=1
651   v[v<=0]=0.0001
652   v[v>1]=1
653   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
654   #out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
655   out <- (u^(-1))*exp(-((-log(u))^theta+(-log(v))^theta)^(1/theta))
656   return(out)
657 }
658
659 # Función a maximizar
660 f_aux <- function(para){
661   D_u_0<-Du_Gum(u,v,para)-Du_Gum(u,vv,para)
662   D_u_0[D_u_0<=0]=1
663   D_u_0 <- log(D_u_0)
664   out<-(sum(D_u_0))
665   return(out)
666 }
667
668 # Proceso de maximización
669 #para_ini <- theta2z(theta_ini,family = family)
670 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
671 #theta_0 <- z2theta(para_ifm,family = family)
672 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
  para la regresión conjunta
673 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
674
675 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
676
677 # Creamos nuestra función a maximizar
678 ##########
679 f_aux_reg <- function(para){
680
681   p <- ncol(var_x)
682   q <- ncol(var_x)
683   alpha <- para[1:p]
684   beta <- para[(p+1):(p+q)]
685   # theta <- z2theta(para[p+q+1],family)

```

```

686     theta <- z2theta(para[p+q+1], family = family)
687     delta <- para[p+q+2]
688     lambda <- as.vector(exp(var_y%%beta)*data_model.Expuesto)
689     mu <- as.vector(exp(var_y%%alpha))
690     #dummy <- density_joint(x,y,mu,delta,lambda,family=family ,zt=FALSE)
691     u2 <- pgam(data_model.Incurrido ,mu, delta)
692     v2<-ppois(y,lambda)
693     vv2<-ppois(y-1,lambda)
694
695     marginal.x <- dgam(data_model.Incurrido ,mu, delta)
696     marginal.x[marginal.x>=1]=0
697     marginal.x[marginal.x<=0]=0
698     marginal.x[is.na(marginal.x)]=0
699     marginal.x[marginal.x==Inf]=0
700     marginal.x[marginal.x== -Inf]=0
701
702     par_der <- Du_Clay(u2,v2,theta)
703     par_der1 <- Du_Clay(u2,vv2,theta)
704
705     dummy<- par_der-par_der1
706
707     dummy[y==0]=par_der[y==0]
708
709     out<-marginal.x*dummy
710
711     out[out<=0]=1e-10
712     ll <- -sum(log(out))
713
714     #if(negative==TRUE) ll <- (-ll)
715     return(ll)
716 }
#####
717 para_0<-c(alpha_0,beta_0,theta_0,family=family),delta_0) # Parámetros iniciales
718 a<-now()
719 para_optim<- optim(para_0,f_aux_reg,method = "BFGS",hessian = FALSE) # Estimación Maxima verosímil de los
720     #parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
721 b<-now()
722 b-a #3.485547 hours
723 #load("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
724     Parcial/JointModel_Gaus.RData")
725 para_optim.value # Log_Likelihood
726 vector <- para_optim.par
727 p <- ncol(var_y)
728 q <- ncol(var_y)
729 Log_Likelihood<-para_optim.value
730 # Parámetros optimizados
731 alpha_gum<-vector[1:p]
732 beta_gum<-vector[(p+1):(p+q)]
733 theta_gum=z2theta(vector[p+q+1],family) #-0.09506928
734 delta_gum<-exp(vector[p+q+2]) #4.271691
735 tau_gum<-BiCopPar2Tau(par=theta_gum,family=family)#-0.06061453
736 head(beta_gum)
737 head(alpha_gum)
738
739 # Determinamos los errores standar para el test de Wald
740 hessian_gum <- para_optim.hessian
741 Hinv <- ginv(hessian_gum)
742 sd <- sqrt(diag(Hinv))
743 sd.alpha_gum <-sd[1:p]
744 sd.beta_gum <- sd[(p+1):(p+q)]
745 sd.theta_gum <-sd[p+q+1]
746
747 # Ordenamos los coeficientes estimados en una tabla
748 Reg_Cop_Gausiana <- data.frame(Estimate_Frec=round(beta_gau,4),Std.error.frecc=round(sd.beta_gau,4),z_
749     _value_frec=round(beta_gau/sd.beta_gau,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_gau/sd.beta_gau))),4),
750     Estimate_Sev=round(alpha_gau,4),Std.error.sev=round(sd.alpha_gau,4),z_value_sev=round(alpha_gau/sd.alpha_
751     _gau,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_gau/sd.alpha_gau))),4))
752 Reg_Cop_Gumbel <- data.frame(Estimate_Frec=round(beta_gum,4),Std.error.frecc=0,z_value_frec=0,"Prob_frec"
753     =0,Estimate_Sev=round(alpha_gum,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
754 var_level<-row.names(Reg_Cop_Gumbel)
755 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
756
757 # Agregamos los interceptos
758 temp <- tibble(
759     var_level = c(
760         "Año2022",

```

```

756 "PP_Tipo_PolizaIndividual",
757 "PP_AntiguedadC_02",
758 "PP_Canal_VentaCorredores",
759 "PP_Uso_VehParticular",
760 "PP_Marca_VehL3",
761 "PP_ZonaA2",
762 "PP_GeneroFemenino",
763 "PP_Edad38-42",
764 "PP_SAj.[13000-999999]"
765 ),
766 Estimate_Frec = 0
767 ,
768 Estimate_Sev = 0
769 )
770
771 # coeficientes del modelo de regresion con dependencia
772 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)` = round(exp(Estimate_Frec),4), `exp(Estimate_Sev)` = round(exp(Estimate_Sev),4)))
773 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% relocate(`exp(Estimate_Frec)` , .after=Prob_frec)
774 # Guardamos el modelo conjunto y los coeficientes
775 write.csv(Reg_Cop_Gumbel, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/Coeff_Gumb.csv", fileEncoding = "Latin1")
776 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/JointModel_Gumb.RData")
777
778 # Simulacion de la Pérdida Total
779 # =====
780
781 # CON DEPENDENCIA
782 # -----
783 lambda_loss_dep <- exp(var_x%*%beta_clay)*data_CM.Expuesto
784 mu_loss_dep <- exp(var_x%*%alpha_clay)
785 delta_loss_dep <- delta_clay
786 # Función de densidad fx
787
788 f_loss <-
789   function(loss ,mu,d,lambda){
790     n<-length(loss)
791     if (length(lambda)==1) lambda <- rep(lambda,n)
792     if (length(mu)==1) mu <- rep(mu,n)
793     out <- vector(length=n)
794
795     for(i in 1:n){
796       N <- 1:20
797       v <- rpois(N,lambda[i])
798       vv <- rpois(N-1,lambda[i])
799       u <- rgam(loss[i]/N,mu[i],d)
800
801       Der_cop <- Du_Clay(u,v,theta)-Du_Clay(u,vv,theta)
802       dummy <- Der_cop*dgam(loss[i]/N,mu[i],d)/N
803       out[i] <- sum(dummy)
804     }
805
806     out[loss <=0]=0
807     return(out)
808   }
809
810 # Función de densidad acumulada Fx
811 F_loss <-
812   function(loss ,mu,d,lambda){
813     out<-vector(length=length(loss))
814
815     for(i in 1:length(loss)){
816       floss <- function(s){
817         f_loss(s,mu[i],d,lambda)
818       }
819       out[i] <- integrate(floss ,0 ,loss[i]).value
820     }
821     return(out)
822   }
823
824 # Estimación del Total Loss (fuerte dedicacion computacional)
825
826 #k <- length(lambda_loss_dep)
827 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
828 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)

```

```

829 L <- vector(length=k)
830 S <- vector(length=m)
831
832 for(j in 1:m){
833   r_uni<-runif(k)
834   for(i in 1:k){
835     f_root <-function(s){
836       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_clay)-r_uni[i]
837     }
838
839     tryCatch(
840       error = function(cond) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
841       loss<-uniroot(f_root,lower = 0,upper = 500000).root
842     )
843
844     #print(loss.root)
845     L[i]<-loss
846     perc<-paste(i/k*100,"%")
847     print(perc)
848   }
849   S[j]<-sum(L)
850   perc<-paste(j/m*100,"%")
851   print(perc)
852   #print(S[j])
853 }
854 S
855
856 # CON INDEPENDENCIA
857 # -----
858 lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO
859 mu_loss <-predict(glm.Sev.PP, newdata=data_CM, type='response')
860 delta_loss <-delta_0
861
862 M<-seq(1:10000)
863 for(i in 1:10000){
864   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
865   S<-vector(length=length(lambda_loss))
866   for(j in 1:length(muestra_cant)){
867     if(muestra_cant[j]==0){
868       S[j]=0
869     } else {
870       S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
871     }
872   }
873
874   M[i] <- sum(S)
875   porc<-paste((i/10000)*100,"%")
876   print(porc)
877   #print(M[i])
878 }
879
880 hist(M)
881
882 # Simulacion de la Frecuencia con Independencia
883 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
884 # Frecuencia
885 N<-seq(1:10000)
886 for(i in 1:10000){
887   N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
888   porc<-paste((i/10000)*100,"%")
889   print(porc)
890 }
891 hist(N)
892
893 ######
894 # Limpiamos el ambiente de variables
895 rm(list = ls())
896 gc()
897
898 ######
899 # REGRESION FRANK
900 #####
901
902 # Set Directory
903 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_"

```

```

904      Parcial")
905
906      # Carga los modelos GLM MARGINALES calculados en otros querys
907      load("glm.Frec.PP.RData")
908      load("glm.Sev.PP.RData")
909      load("Data_PP.RData")
910      load("Data_PP_model.RData")
911      load("Data_PP_model_CM.RData")
912
913      # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
914      # SEVERIDAD
915      x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
916      var_x <- model.matrix(glm.Sev.PP) #data_CM[,3:12]
917      # FRECUENCIA
918      y <- data_model.Cantidad
919      Expuestos <- data_model.Expuesto
920      var_y <- model.matrix(glm.Frec.PP) #data[,3:12]
921
922      # Estimacion de parámetros iniciales utilizando las distribuciones marginales
923      # SEVERIDAD
924      sd_alpha <- sqrt(diag(vcov(glm.Sev.PP)))
925      alpha_0 <- glm.Sev.PP.coefficients
926      delta_0 <- summary(glm.Sev.PP).dispersion
927      mu_0 <- exp(var_x%*%alpha_0)
928
929      # FRECUENCIA
930      beta_0 <- glm.Frec.PP.coefficients
931      sd_beta <- sqrt(diag(vcov(glm.Frec.PP)))
932      lambda_0 <- exp(var_y%*%beta_0)*Expuestos
933
934      # REGRESION MEDIANTE COPULAS
935
936      # COPULA GUMBEL
937      family=5
938      # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
939      theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
940      lambda_0)/(length(y)+1), family=family).par
941      #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
942      data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
943      tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
944
945      # Estimación de las seudo-observaciones
946      u<-pgam(x,mu_0,delta_0)
947      v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
948      vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
949
950      # Creamos la función primera derivada de la Función Copula
951      Du_Frank <- function(u,v,theta){
952          u[u<=0]=0.0001
953          u[u>1]=1
954          v[v<=0]=0.0001
955          v[v>1]=1
956          #out <- pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
957          #out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta)-1)
958          #out <- (u^(-1))*exp(-((-log(u))^theta+(-log(v))^theta)^(1/theta))
959          out <- (exp(theta)*(exp(theta*v)-1))/(exp(theta*(u+1))+exp(theta*(v+1))-exp(theta)-exp(theta*(u+v)))
960          return(out)
961      }
962
963      # Función a maximizar
964      f_aux <- function(para){
965          D_u_0<-Du_Frank(u,v,para)-Du_Frank(u,vv,para)
966          D_u_0[D_u_0<=0]=1
967          D_u_0 <- log(D_u_0)
968          out<-(-sum(D_u_0))
969          return(out)
970      }
971
972      # Proceso de maximización
973      #para_ini <- theta2z(theta_ini,family = family)
974      para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
975      #theta_0 <- z2theta(para_ifm,family = family)
976      theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
977      # para la regresión conjunta
978      tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia

```

```

976 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
977
978 # Creamos nuestra función a maximizar
979 ##########
980 f_aux_reg <- function(para){
981
982   p      <- ncol(var_x)
983   q      <- ncol(var_x)
984   alpha  <- para[1:p]
985   beta   <- para[(p+1):(p+q)]
986   # theta <- z2theta(para[p+q+1],family)
987   theta  <- z2theta(para[p+q+1],family = family)
988   delta   <- para[p+q+2]
989   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
990   mu     <- as.vector(exp(var_y%*%alpha))
991   #dummy  <- density_joint(x,y,mu,delta,lambda,theta,family=family,zt=FALSE)
992   u2 <- pgam(data_model.Incurrido,mu,delta)
993   v2<-ppois(y,lambda)
994   vv2<-ppois(y-1,lambda)
995
996   marginal.x <- dgam(data_model.Incurrido,mu,delta)
997   marginal.x[marginal.x>=1]=0
998   marginal.x[marginal.x<=0]=0
999   marginal.x[is.na(marginal.x)]=0
1000  marginal.x[marginal.x== -Inf]=0
1001  marginal.x[marginal.x== Inf]=0
1002
1003  par_der  <- Du_Frank(u2,v2,theta)
1004  par_der1 <- Du_Frank(u2,vv2,theta)
1005
1006  dummy<- par_der-par_der1
1007
1008  dummy[y==0]=par_der[y==0]
1009
1010  out<-marginal.x*dummy
1011
1012  out[out<=0]=1e-10
1013  ll       <- -sum(log(out))
1014
1015  #if(negative==TRUE) ll <- (-ll)
1016  return(ll)
1017 }
1018
1019 ##########
1020 para_0<-c(alpha_0,beta_0,theta_0,family=family),delta_0) # Parámetros iniciales
1021 a<-now()
1022 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = FALSE) # Estimación Maxima verosmil de los
1023 parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
1024 b<-now()
1025 b-a #2.82135 hours
1026 #load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
1027 _Parcial/JointModel_Gaus.RData")
1028 para_optim.value # Log_Likelihood
1029 vector <- para_optim.par
1030 p      <- ncol(var_y)
1031 q      <- ncol(var_y)
1032 Log_Likelihood<-para_optim.value
1033 # Parámetros optimizados
1034 alpha_frank<-vector[1:p]
1035 beta_frank<-vector[(p+1):(p+q)]
1036 theta_frank=z2theta(vector[p+q+1],family)
1037 delta_frank<-exp(vector[p+q+2]) #4.271691
1038 tau_frank<-BiCopPar2Tau(par=theta_frank,family=family)
1039 head(beta_frank)
1040 head(alpha_frank)
1041
1042 # Determinamos los errores standar para el test de Wald
1043 hessian_frank <- para_optim.hessian
1044 Hinvt      <- ginv(hessian_frank)
1045 sd          <- sqrt(diag(Hinvt))
1046 sd.alpha_frank <-sd[1:p]
1047 sd.beta_frank <- sd[(p+1):(p+q)]
1048 sd.theta_frank <-sd[p+q+1]
1049
1050 # Ordenamos los coeficientes estimados en una tabla
1051 #Reg_Cop_Gausiana <- data.frame(Estimate_Freq=round(beta_gau,4),Std.error.freq=round(sd.beta_gau,4),z_

```

```

1050   value_frec=round(beta_gau/sd.beta_gau,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_gau/sd.beta_gau))),4),
1051   Estimate_Sev=round(alpha_gau,4),Std.error.sev=round(sd.alpha_gau,4),z_value_sev=round(alpha_gau/sd.alpha_gau,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_gau/sd.alpha_gau))),4))
1052 Reg_Cop_Frank <- data.frame(Estimate_Frec=round(beta_frank,4),Std.error.freq=0,z_value_frec=0,"Prob_frec"
1053   =0,Estimate_Sev=round(alpha_frank,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
1054 var_level<-row.names(Reg_Cop_Frank)
1055 Reg_Cop_Frank <- Reg_Cop_Frank %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
1056
1057 # Agregamos los interceptos
1058 temp <- tibble(
1059   var_level = c(
1060     "Año2022",
1061     "PP_Tipo_PolizaIndividual",
1062     "PP_Antiguedad_c_02",
1063     "PP_Canal_VentaCorredores",
1064     "PP_Uso_VehParticular",
1065     "PP_Marca_VehL3",
1066     "PP_ZonaA2",
1067     "PP_GeneroFemenino",
1068     "PP_Edad38-42",
1069     "PP_SAj.[13000-999999]"
1070   ),
1071   Estimate_Frec = 0
1072   ,
1073   Estimate_Sev = 0
1074 )
1075
1076 # coeficientes del modelo de regresion con dependencia
1077 Reg_Cop_Frank <- Reg_Cop_Frank %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)`=
1078   round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
1079 Reg_Cop_Frank <-Reg_Cop_Frank %>% relocate(`exp(Estimate_Frec)`, .after=Prob_frec)
1080
1081 # Guardamos el modelo conjunto y los coeficientes
1082 write.csv(Reg_Cop_Frank, file="C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/Coef_Frank.csv", fileEncoding = "Latin1")
1083 save(para_optim, file="C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/JointModel_Frank.RData")
1084
1085 # Simulacion de la Perdida Total
1086 # =====
1087
1088 # CON DEPENDENCIA
1089
1090 lambda_loss_dep <- exp(var_x%*%beta_frank)*data_CM.Expuesto
1091 mu_loss_dep <- exp(var_x%*%alpha_frank)
1092 delta_loss_dep <- delta_frank
1093
1094 # Función de densidad fx
1095 f_loss <-
1096   function(loss ,mu,d, lam ,theta){
1097     n<-length(loss)
1098     if (length(lam)==1) lam <- rep(lam,n)
1099     if (length(mu)==1) mu <- rep(mu,n)
1100     out <- vector(length=n)
1101
1102     for(i in 1:n){
1103       N <-1:20
1104       v <-ppois(N, lam[i])
1105       vv <-ppois(N-1, lam[i])
1106       u <-pgam(loss[i]/N,mu[i],d)
1107
1108       Der_cop <-Du_Frank(u,v,theta)-Du_Frank(u,vv,theta) # Primera derivada de la función copula
1109       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N # Teorela del Total Loss
1110       out[i] <-sum(dummy)
1111     }
1112
1113     out[loss <=0]=0
1114     return(out)
1115   }
1116
1117 # Función de densidad acumulada Fx
1118 F_loss <-
1119   function(loss ,mu,d, lam ,theta){
1120     out<-vector(length=length(loss)))

```

```

1119
1120     for(i in 1:length(loss)){
1121         floss <- function(s){
1122             f_loss(s,mu[i],d,lambda[i],theta)
1123         }
1124         out[i] <- integrate(floss ,0 ,loss[i]).value # Integral para el cálculo del a F(x) a partir de f(x)
1125     }
1126     return(out)
1127 }
1128
1129 # Estimación del Total Loss (fuerte dedicacion computacional)
1130
1131 #k <-length(lambda_loss_dep)
1132 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
1133 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
1134 L <- vector(length=k)
1135 S <- vector(length=m)
1136
1137 for(j in 1:m){
1138     r_uni<-runif(k)
1139     for(i in 1:k){
1140         f_root <-function(s){
1141             F_loss(s,mu_loss_dep[i],delta_loss_dep ,lambda_loss_dep[i],theta_frank)-r_uni[i]
1142         }
1143
1144         tryCatch(
1145             error = function(cnd) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
1146             loss<-uniroot(f_root ,lower = 0 ,upper = 500000).root
1147         )
1148
1149         #print(loss.root)
1150         L[i]<-loss
1151         perc<-paste(i/k*100,"%")
1152         print(perc)
1153     }
1154     S[j]<-sum(L)
1155     perc<-paste(j/m*100,"%")
1156     print(perc)
1157     #print(S[j])
1158 }
1159 S
1160
1161 # CON INDEPENDENCIA
1162 # -----
1163 lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO
1164 mu_loss <-predict(glm.Sev.PP, newdata=data_CM, type='response')
1165 delta_loss <-delta_0
1166
1167 M<-seq(1:10000)
1168 for(i in 1:10000){
1169     muestra_cant<-rpois(length(lambda_loss),lambda_loss)
1170     S<-vector(length=length(lambda_loss))
1171     for(j in 1:length(muestra_cant)){
1172         if(muestra_cant[j]==0){
1173             S[j]=0
1174         } else{
1175             S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
1176         }
1177     }
1178
1179     M[i] <- sum(S)
1180     porc<-paste((i/10000)*100,"%")
1181     print(porc)
1182     #print(M[i])
1183 }
1184
1185 hist(M)
1186
1187 # Simulacion de la Frecuencia con Independencia
1188 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
1189 # Frecuencia
1190 N<-seq(1:10000)
1191 for(i in 1:10000){
1192     N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
1193     porc<-paste((i/10000)*100,"%")

```

```

1194     print(porc)
1195   }
1196   hist(N)

```

C.4. Modelo Conjunto - Pérdida Total

```

1 ##########
2 ##########
3 #####
4 ##### REGRESION BASADO EN COPULAS #####
5 ##### PERDIDA TOTAL #####
6 #####
7 ##########
8 ##########
9
10 # Cargamos las librerías necesarias
11
12 #library(copula)
13 library(RODBC)
14 library(tidyverse)
15 library(MASS)
16 #library(GJRM)
17 #library(devtools)
18 library(CopulaRegression)
19 library(VineCopula)
20 #library(optimx)
21 #install_url('http://cran.r-project.org/src/contrib/Archive/CopulaRegression/CopulaRegression_0.1-5.tar.gz')
22
23 ##########
24 # REGRESION GAUSSIANA
25 ##########
26
27 # Set Directory
28 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Total")
29
30 # Carga los modelos GLM MARGINALES calculados en otros querys
31 load("glm.Frec.RData")
32 load("glm.Sev.RData")
33 load("Data_PT.RData")
34 load("Data_PT_model.RData")
35 load("Data_PT_model_CM.RData")
36
37 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
38 # SEVERIDAD
39 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
40 var_x <- model.matrix(glm.Sev.PT) #data_CM[,3:12]
41 # FRECUENCIA
42 y <- data_model.Cantidad
43 Expuestos <- data_model.Expuesto
44 var_y <- model.matrix(glm.Frec.PT) #data[,3:12]
45
46 # Estimacion de parámetros iniciales utilizando las distribuciones marginales
47 # SEVERIDAD
48 sd_alpha <- sqrt(diag(vcov(glm.Sev.PT)))
49 alpha_0 <- glm.Sev.PT.coefficients
50 delta_0 <- summary(glm.Sev.PT).dispersion
51 mu_0 <- exp(var_x%*%alpha_0)
52
53 # FRECUENCIA
54 beta_0 <- glm.Frec.PT.coefficients
55 sd_beta <- sqrt(diag(vcov(glm.Frec.PT)))
56 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
57
58 # REGRESION MEDIANTE COPULAS
59
60 # COPULA GAUSSIANA
61 family=1
62 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
63 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-lambda_0)/(length(y)+1), family=family).par
64 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
65 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
66

```

```

67 # Estimación de las seudo-observaciones
68 u<-pgam(x,mu_0,delta_0)
69 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%%beta_0)*data_CM.Expuesto)
70 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%%beta_0)*data_CM.Expuesto)
71
72 # Creamos la función primera derivada de la Función Copula
73 Du_Gaus <- function(u,v,theta){
74   u[u<=0]=0.001
75   u[u>=1]=0.999
76   v[v<=0]=0.001
77   v[v>=1]=0.999
78   out <- pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
79   return(out)
80 }
81
82 # Función a maximizar
83 f_aux <- function(para){
84   D_u_0<-Du_Gaus(u,v,para)-Du_Gaus(u,vv,para)
85   D_u_0[D_u_0<=0]=1
86   D_u_0 <- log(D_u_0)
87   out<-(-sum(D_u_0))
88   return(out)
89 }
90
91 # Proceso de maximización
92 #para_ini <- theta2z(theta_ini,family = family)
93 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
94 #theta_0 <- z2theta(para_ifm,family = family)
95 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
# para la regresión conjunta
96 tau <- BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
97
98 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
99
100 # Creamos nuestra función a maximizar
101 ##########
102 f_aux_reg <- function(para){
103
104   p <- ncol(var_x)
105   q <- ncol(var_x)
106   alpha <- para[1:p]
107   beta <- para[(p+1):(p+q)]
108   # theta <- z2theta(para[p+q+1],family)
109   theta <- para[p+q+1]
110   delta <- para[p+q+2]
111   lambda <- as.vector(exp(var_y%%beta)*data_model.Expuesto)
112   mu <- as.vector(exp(var_y%%alpha))
113   #dummy <- density_joint(x,y,mu,delta,theta,family=family,zt=FALSE)
114   u2 <- pgam(data_model.Incurrido,mu,delta)
115   v2<-ppois(y,lambda)
116   vv2<-ppois(y-1,lambda)
117
118   marginal.x <- dgam(data_model.Incurrido,mu,delta)
119   marginal.x[marginal.x>=1]=0
120   marginal.x[marginal.x<=0]=0
121   marginal.x[is.na(marginal.x)]=0
122   marginal.x[marginal.x==Inf]=0
123   marginal.x[marginal.x==Inf]=0
124
125   par_der <- Du_Gaus(u2,v2,theta)
126   par_der1 <- Du_Gaus(u2,vv2,theta)
127
128   dummy<- par_der-par_der1
129
130   dummy[y==0]=par_der[y==0]
131
132   out<-marginal.x*dummy
133
134   out[out<=0]=1e-10
135   ll <- -sum(log(out))
136
137   #if(negative==TRUE) ll <- (-ll)
138   return(ll)
139 }
140
141 #####

```

```

142
143 para_0<-c(alpha_0,beta_0,theta_0,delta_0) # Parámetros iniciales
144 a<-now()
145 para_optim <- optim(para_0,f_aux_reg ,method = "BFGS",hessian = TRUE) # Demora 6min
146 b<-now()
147 b-a
148 #load ("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
149 Parcial/JointModel_Gaus.RData")
150 para_optim.value # Log_Likelihood
151 vector <- para_optim.par
152 p <- ncol(var_y)
153 q <- ncol(var_y)
154
155 # Parámetros optimizados
156 alpha_gau<-vector[1:p]
157 beta_gau<-vector[(p+1):(p+q)]
158 theta_gau=z2theta(vector[p+q+1],family) # -0.09506928
159 delta_gau<-exp(vector[p+q+2]) # 4.271691
160 tau_gau<-BiCopPar2Tau(par=theta_gau,family=family) #-0.06061453
161 head(beta_gau)
162 head(alpha_gau)
163 # Determinamos los errores estandar para el test de Wald
164 hessian_gau <- para_optim.hessian
165 Hinv <- ginv(hessian_gau)
166 sd <- sqrt(diag(Hinv))
167 sd.alpha_gau <-sd[1:p]
168 sd.beta_gau <- sd[(p+1):(p+q)]
169 sd.theta_gau <-sd[p+q+1]
170
171 # Ordenamos los coeficientes estimados en una tabla
172 Reg_Cop_Gausiana <- data.frame(Estimate_Frec=round(beta_gau,4),Std.error.freq=round(sd.beta_gau,4),z_
173 value.freq=round(beta_gau/sd.beta_gau,4),"Prob.freq"=round(2*(1-pnorm(abs(beta_gau/sd.beta_gau))),4),
174 Estimate_Sev=round(alpha_gau,4),Std.error.sev=round(sd.alpha_gau,4),z_value_sev=round(alpha_gau/sd.alpha_
175 gau,4),"Prob.Sev"=round(2*(1-pnorm(abs(alpha_gau/sd.alpha_gau))),4))
176
177 # Agregamos los interceptos
178 temp <- tibble(
179   var_level = c(
180     "Año2022",
181     "PT_Tipo_PolizaIndividual",
182     "PT_Uso_VehParticular",
183     "PT_Marca_VehL4",
184     "PT_ZonaA3",
185     "PT_GeneroFemenino",
186     "PT_Edad43-52"
187   ),
188   Estimate_Frec = 0
189   ,
190   Estimate_Sev = 0
191 )
192
193 # coeficientes del modelo de regresion con dependencia
194 Reg_Cop_Gausiana <- Reg_Cop_Gausiana %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_
195 _Frec)`=round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
196 Reg_Cop_Gausiana <-Reg_Cop_Gausiana %>% relocate(`exp(Estimate_Frec)`,.after=Prob.freq)
197
198 # Guardamos el modelo conjunto y los coeficientes
199 write.csv(Reg_Cop_Gausiana , file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
200 GLM_Conjuntos/Perdida_Total/Coef_Gaus.csv", fileEncoding = "Latin1")
201 save(para_optim , file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
202 Conjuntos/Perdida_Total/JointModel_Gaus.RData")
203
204 # Simulacion de la Pérdida Total
205 # =====
206
207 # CON DEPENDENCIA
208 # -----
209 lambda_loss_dep <- exp(var_x%*%beta_gau)*data_CM.Expuesto
210 mu_loss_dep <- exp(var_x%*%alpha_gau)
211 delta_loss_dep <- delta_gau
212 # Función de densidad fx
213 f_loss <-
214   function(loss,mu,d,lam,theta){
215     n<-length(loss)
216     if (length(lam)==1) lam <- rep(lam,n)
217     if (length(mu)==1) mu <- rep(mu,n)
218   }

```

```

211          out           <- vector(length=n)
212
213      for(i in 1:n){
214          N  <-1:20
215          v  <-rpois(N, lam[i])
216          vv <-rpois(N-1, lam[i])
217          u  <-pgam(loss[i]/N,mu[i],d)
218
219          Der_cop <-Du_Gaus(u,v,theta)-Du_Gaus(u,vv,theta)
220          dummy   <-Der_cop*dgam(loss[i]/N,mu[i],d)/N
221          out[i]  <-sum(dummy)
222      }
223
224      out[loss<=0]=0
225      return(out)
226  }
227
228 # Función de densidad acumulada Fx
229 F_loss <-
230 function(loss ,mu,d, lam ,theta){
231     out<-vector(length=length(loss))
232
233     for(i in 1:length(loss)){
234         floss <- function(s){
235             f_loss(s,mu[i],d, lam[i], theta)
236         }
237         out[i] <- integrate(floss ,0 ,loss[i]).value
238     }
239     return(out)
240 }
241
242 # Estimación del Total Loss (fuerte dedicacion computacional)
243
244 #k <-length(lambda_loss_dep)
245 k <- length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
246 m <- 100 # Cantidad de muestras aleatorias (Deben ser 10 mil)
247 L <- vector(length=k)
248 S <- vector(length=m)
249
250 for(j in 1:m){
251     r_uni<-runif(k)
252     for(i in 1:k){
253         f_root <-function(s){
254             F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_gau)-r_uni[i]
255         }
256
257         tryCatch(
258             error = function(cond) mu_loss_dep[i]*lambda_loss_dep[i],
259             loss<-uniroot(f_root ,lower = 0,upper = 500000)
260         )
261
262         #print(loss.root)
263         L[i]<-loss.root
264         perc<-paste(i/k*100,"%")
265         print(perc)
266     }
267     S[j]<-sum(L)
268     perc<-paste(j/m*100,"%")
269     print(perc)
270     #print(S[j])
271 }
272 head(S)
273
274 # CON INDEPENDENCIA
275 #
276 lambda_loss <-predict(glm.Frec.PT, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN
277 MODELO DE RIESGO COLECTIVO
278 mu_loss    <-predict(glm.Sev.PT, newdata=data_CM, type='response')
279 delta_loss <-delta_0
280
281 M<-seq(1:10000)
282 for(i in 1:10000){
283     muestra_cant<-rpois(length(lambda_loss),lambda_loss)
284     S<-vector(length=length(lambda_loss))
285     for(j in 1:length(muestra_cant)){
286         if(muestra_cant[j]==0){

```

```

286           S[j]=0
287       } else {
288           S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
289       }
290   }
291
292   M[i] <- sum(S)
293   porc<-paste((i/10000)*100,"%")
294   print(porc)
295   #print(M[i])
296   }
297
298   hist(M,breaks = 50)
299   mean(M)
300   sum(data.Incurrido)
301
302
303 # Simulacion de la Frecuencia con Independencia
304 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
305 # Frecuencia
306 N<-seq(1:10000)
307 for(i in 1:10000){
308   N[i] <- sum(simulate(glm.Frec.PT,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
309   pore<-paste((i/10000)*100,"%")
310   print(pore)
311   }
312   hist(N,breaks = 50)
313   mean(N)
314   mean(data.Frecuencia)
315
316 #####
317 # Limpiamos el ambiente de variables
318 rm(list = ls())
319 gc()
320
321 #####
322 # REGRESION CLAYTON
323 #####
324
325 # Set Directory
326 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_Total")
327
328 # Carga los modelos GLM MARGINALES calculados en otros querys
329 load("glm.Frec.PT.RData")
330 load("glm.Sev.PT.RData")
331 load("Data_PT.RData")
332 load("Data_PT_model.RData")
333 load("Data_PT_model_CM.RData")
334
335 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
336 # SEVERIDAD
337 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
338 var_x <-model.matrix(glm.Sev.PT) #data_CM[,3:12]
339 # FRECUENCIA
340 y <- data_model.Cantidad
341 Expuestos <- data_model.Expuesto
342 var_y <- model.matrix(glm.Frec.PT) #data[,3:12]
343
344 # Estimacion de parametros iniciales utilizando las distribuciones marginales
345 # SEVERIDAD
346 sd_alpha <- sqrt(diag(vcov(glm.Sev.PT)))
347 alpha_0 <- glm.Sev.PT.coefficients
348 delta_0 <- summary(glm.Sev.PT).dispersion
349 mu_0 <- exp(var_x%>%alpha_0)
350
351 # FRECUENCIA
352 beta_0 <- glm.Frec.PT.coefficients
353 sd_beta <- sqrt(diag(vcov(glm.Frec.PT)))
354 lambda_0 <- exp(var_y%>%beta_0)*Expuestos
355
356 # REGRESION MEDIANTE COPULAS
357
358 # COPULA CLAYTON
359 family=3
360 # Estimacion inicial del parametro theta para estimarlo mediante el algoritmo de MPLE

```

```

361 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%%alpha_0))/(length(y)+1),rank(y*Expuestos-
362 lambda_0)/(length(y)+1), family=family).par
#theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
363 data_CM.Expuesto-exp(var_x%%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)

365 # Estimación de las seudo-observaciones
366 u<-pgam(x,mu_0,delta_0)
367 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%%beta_0)*data_CM.Expuesto)
368 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%%beta_0)*data_CM.Expuesto)
369

370 # Creamos la función primera derivada de la Función Copula
371 Du_Clay <- function(u,v,theta){
372   u[u<=0]=0.0001
373   u[u>1]=1
374   v[v<=0]=0.0001
375   v[v>1]=1
376   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
377   out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
378   return(out)
379 }
380

381 # Función a maximizar
382 f_aux <- function(para){
383   D_u_0<-Du_Clay(u,v,para)-Du_Clay(u,vv,para)
384   D_u_0[D_u_0<=0]=1
385   D_u_0 <- log(D_u_0)
386   out<-(sum(D_u_0))
387   return(out)
388 }
389

390 # Proceso de maximización
391 #para_ini <- theta2z(theta_ini,family = family)
392 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
#theta_0 <- z2theta(para_ifm,family = family)
393 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
# para la regresión conjunta
394 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia

395 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
396
397 # Creamos nuestra función a maximizar
398 #####f_aux_reg <- function(para){

399 f_aux_reg <- function(para){

400   p <- ncol(var_x)
401   q <- ncol(var_x)
402   alpha <- para[1:p]
403   beta <- para[(p+1):(p+q)]
404   # theta <- z2theta(para[p+q+1],family)
405   theta <- z2theta(para[p+q+1],family = family)
406   delta <- para[p+q+2]
407   lambda <- as.vector(exp(var_y%%beta)*data_model.Expuesto)
408   mu <- as.vector(exp(var_y%%alpha))
409   #dummy <- density_joint(x,y,mu,delta,lambda,theta,family=family,zt=FALSE)
410   u2 <- pgam(data_model.Incurrido,mu,delta)
411   v2<-ppois(y,lambda)
412   vv2<-ppois(y-1,lambda)
413

414   marginal.x <- dgam(data_model.Incurrido ,mu, delta)
415   marginal.x[marginal.x>=1]=0
416   marginal.x[marginal.x<=0]=0
417   marginal.x[is.na(marginal.x)]=0
418   marginal.x[marginal.x==Inf]=0
419   marginal.x[marginal.x== -Inf]=0
420

421   par_der <- Du_Clay(u2,v2,theta)
422   par_der1 <- Du_Clay(u2,vv2,theta)
423

424   dummy<- par_der-par_der1
425   dummy[y==0]=par_der[y==0]
426
427   out<-marginal.x*dummy
428
429   out[out<=0]=1e-10
430
431
432
433

```

```

434     11      <- -sum(log(out))
435
436     #if(negative==TRUE) 11 <- (-11)
437     return(11)
438   }
439
440 #####parámetros iniciales
441 para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
442 a<-now()
443 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Estimación Maxima verosímil de los
444     # parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
445 b<-now()
446 b-a
447 #load("C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
448     Parcial/JointModel_Gaus.RData")
449 para_optim.value # Log_Likelihood
450 vector <- para_optim.par
451 p      <- ncol(var_y)
452 q      <- ncol(var_y)
453 Log_Likelihood<-para_optim.value
454 # Parámetros optimizados
455 alpha_clay<-vector[1:p]
456 beta_clay<-vector[(p+1):(p+q)]
457 theta_clay=z2theta(vector[p+q+1],family) # -0.09506928
458 delta_clay<-exp(vector[p+q+2]) #4.271691
459 tau_clay<-BiCopPar2Tau(par=theta_clay,family=family) #-0.06061453
460 head(beta_0)
461 head(beta_clay)
462 head(alpha_0)
463 head(alpha_clay)
464 # Determinamos los errores standar para el test de Wald
465 hessian_clay <- para_optim.hessian
466 Hinv         <- ginv(hessian_clay)
467 sd           <- sqrt(diag(Hinv))
468 sd.alpha_clay <-sd[1:p]
469 sd.beta_clay <- sd[(p+1):(p+q)]
470 sd.theta_clay <-sd[p+q+1]
471
472 # Ordenamos los coeficientes estimados en una tabla
473 Reg_Cop_Clayton <- data.frame(Estimate_Frec=round(beta_clay,4),Std.error.freq=round(sd.beta_clay,4),z_
474     value_frec=round(beta_clay/sd.beta_clay,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_clay/sd.beta_clay))),4),
475     Estimate_Sev=round(alpha_clay,4),Std.error.sev=round(sd.alpha_clay,4),z_value_sev=round(alpha_clay/sd.
476     alpha_clay,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_clay/sd.alpha_clay))),4))
477
478 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% as_tibble() %>% mutate(var_level=var_level) %>% relocate(var_level
479     )
480
481 # Agregamos los interceptos
482 temp <- tibble(
483     var_level = c(
484         "Año2022",
485         "PT_Tipo_PolizaIndividual",
486         "PT_Uso_VehParticular",
487         "PT_Marca_VehL4",
488         "PT_ZonaA3",
489         "PT_GeneroFemenino",
490         "PT_Edad43–52"
491     ),
492     Estimate_Frec = 0
493     ,
494     Estimate_Sev = 0
495 )
496
497 # coeficientes del modelo de regresión con dependencia
498 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_.
499     Frec)`=round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
500
501 Reg_Cop_Clayton <-Reg_Cop_Clayton %>% relocate(`exp(Estimate_Frec)`,.after=Prob_frec)
502
503 # Guardamos el modelo conjunto y los coeficientes
504 write.csv(Reg_Cop_Clayton, file="C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
505     GLM_Conjuntos/Perdida_Total/Coef_Clay.csv", fileEncoding = "Latin1")
506 save(para_optim, file="C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
507     Conjuntos/Perdida_Total/JointModel_Clay.RData")
508
509 # Simulacion de la Pérdida Total

```

```

501 # =====
502 # CON DEPENDENCIA
503 # -----
504 lambda_loss_dep <- exp(var_x%&%beta_clay)*data_CM_Expuesto
505 mu_loss_dep <- exp(var_x%&%alpha_clay)
506 delta_loss_dep <- delta_clay
507 # Función de densidad fx
508 f_loss <-
509   function(loss ,mu,d, lam ,theta){
510     n<-length(loss)
511     if (length(lam)==1) lam <- rep(lam,n)
512     if (length(mu)==1) mu <- rep(mu,n)
513     out <- vector(length=n)
514
515     for(i in 1:n){
516       N <-1:20
517       v <-rpois(N, lam[i])
518       vv <-rpois(N-1, lam[i])
519       u <-pgam(loss[i]/N,mu[i],d)
520
521       Der_cop <-Du_Clay(u,v, theta )-Du_Clay(u,vv, theta )
522       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N
523       out[i] <-sum(dummy)
524     }
525
526     out[loss <=0]=0
527     return(out)
528   }
529
530 # Función de densidad acumulada Fx
531 F_loss <-
532   function(loss ,mu,d, lam ,theta){
533     out<-vector(length=length(loss))
534
535     for(i in 1:length(loss)){
536       floss <- function(s){
537         f_loss(s,mu[i],d, lam[i], theta )
538       }
539       out[i] <- integrate(floss ,0 ,loss[i]).value
540     }
541     return(out)
542   }
543
544
545 # Estimación del Total Loss (fuerte dedicacion computacional)
546
547 #k <-length(lambda_loss_dep)
548 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
549 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
550 L <- vector(length=k)
551 S <- vector(length=m)
552
553 for(j in 1:m){
554   r_uni<-runif(k)
555   for(i in 1:k){
556     f_root <-function(s){
557       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_clay)-r_uni[i]
558     }
559
560     tryCatch(
561       error = function(cnd) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
562       loss<-uniroot(f_root ,lower = 0,upper = 500000).root
563     )
564
565     #print(loss.root)
566     L[i]<-loss
567     perc<-paste(i/k*100,"%")
568     print(perc)
569   }
570   S[j]<-sum(L)
571   perc<-paste(j/m*100,"%")
572   print(perc)
573   #print(S[j])
574 }
575 S
576

```

```

577 # CON INDEPENDENCIA
578 # -----
579 lambda_loss <- predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO
580 mu_loss <- predict(glm.Sev.PP, newdata=data_CM, type='response')
581 delta_loss <- delta_0
582
583 M<-seq(1:10000)
584 for(i in 1:10000){
585   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
586   S<-vector(length=length(lambda_loss))
587   for(j in 1:length(muestra_cant)){
588     if(muestra_cant[j]==0){
589       S[j]=0
590     } else {
591       S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
592     }
593   }
594
595   M[i] <- sum(S)
596   porc<-paste((i/10000)*100,"%")
597   print(porc)
598   #print(M[i])
599 }
600
601 hist(M)
602
603 # Simulacion de la Frecuencia con Independencia
604 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
605 # Frecuencia
606 N<-seq(1:10000)
607 for(i in 1:10000){
608   N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
609   porc<-paste((i/10000)*100,"%")
610   print(porc)
611 }
612 hist(N)
613
614 ##### Limpiamos el ambiente de variables
615 rm(list = ls())
616 gc()
617
618 #####
619 # REGRESION GUMBEL
620 #####
621 #####
622
623 # Set Directory
624 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_"
Total")
625
626 # Carga los modelos GLM MARGINALES calculados en otros querys
627 load("glm.Frec.PT.RData")
628 load("glm.Sev.PT.RData")
629 load("Data_PT.RData")
630 load("Data_PT_model.RData")
631 load("Data_PT_model_CM.RData")
632
633 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
634 # SEVERIDAD
635 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
636 var_x <- model.matrix(glm.Sev.PT) #data_CM[,3:12]
637 # FRECUENCIA
638 y <- data_model.Cantidad
639 Expuestos <- data_model.Expuesto
640 var_y <- model.matrix(glm.Frec.PT) #data[,3:12]
641
642 # Estimacion de parametros iniciales utilizando las distribuciones marginales
643 # SEVERIDAD
644 sd_alpha <- sqrt(diag(vcov(glm.Sev.PT)))
645 alpha_0 <- glm.Sev.PT.coefficients
646 delta_0 <- summary(glm.Sev.PT).dispersion
647 mu_0 <- exp(var_x%*%alpha_0)
648
649 # FRECUENCIA
650 beta_0 <- glm.Frec.PT.coefficients

```

```

651 sd_beta <- sqrt(diag(vcov(glm.Frec.PT)))
652 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
653
654 # REGRESION MEDIANTE COPULAS
655
656 # COPULA GUMBEL
657 family=4
658 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
659 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
  lambda_0)/(length(y)+1), family=family).par
660 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
  data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
661 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
662
663 # Estimación de las seudo-observaciones
664 u<-pgam(x,mu_0,delta_0)
665 v <- ppois(data_CM.Cantidad ,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
666 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
667
668 # Creamos la función primera derivada de la Función Copula
669 Du_Gum <- function(u,v,theta){
670   u[u<=0]=0.0001
671   u[u>1]=1
672   v[v<=0]=0.0001
673   v[v>1]=1
674   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
675   #out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
676   out <- (u^-1)*exp((-(-log(u))^theta+(-log(v))^theta)^(1/theta))
677   return(out)
678 }
679
680 # Función a maximizar
681 f_aux <- function(para){
682   D_u_0<-Du_Gum(u ,v ,para)-Du_Gum(u ,vv ,para)
683   D_u_0[D_u_0<=0]=1
684   D_u_0 <- log(D_u_0)
685   out<-(sum(D_u_0))
686   return(out)
687 }
688
689 # Proceso de maximización
690 #para_ini <- theta2z(theta_ini ,family = family)
691 para_ifm <- optim(theta_ini ,f_aux,method = "BFGS").par
692 #theta_0 <- z2theta(para_ifm,family = family)
693 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
  para la regresión conjunta
694 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
695
696 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
697
698 # Creamos nuestra función a maximizar
699 #####
700 f_aux_reg <- function(para){
701
702   p <- ncol(var_x)
703   q <- ncol(var_x)
704   alpha <- para[1:p]
705   beta <- para[(p+1):(p+q)]
706   # theta <- z2theta(para[p+q+1],family)
707   theta <- z2theta(para[p+q+1],family = family)
708   delta <- para[p+q+2]
709   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
710   mu <- as.vector(exp(var_y%*%alpha))
711   #dummy <- density_joint(x,y,mu,delta ,lambda ,family=family ,zt=FALSE)
712   u2 <- pgam(data_model.Incurrido ,mu, delta )
713   v2<-ppois(y,lambda)
714   vv2<-ppois(y-1,lambda)
715
716   marginal.x <- dgam(data_model.Incurrido ,mu, delta )
717   marginal.x[marginal.x>=1]=0
718   marginal.x[marginal.x<=0]=0
719   marginal.x[is.na(marginal.x)]=0
720   marginal.x[marginal.x== -Inf]=0
721   marginal.x[marginal.x== Inf]=0
722
723   par_der <- Du_Gum(u2 ,v2 ,theta)

```

```

724     par_der1 <- Du_Gum(u2, vv2, theta)
725
726     dummy<- par_der-par_der1
727
728     dummy[y==0]=par_der[y==0]
729
730     out<-marginal.x*dummy
731
732     out[out<=0]=1e-10
733     ll      <- -sum(log(out))
734
735     #if(negative==TRUE) ll <- (-ll)
736     return(ll)
737 }
738
739 #####para_0<-c(alpha_0,beta_0,theta_0,family=family),delta_0) # Parámetros iniciales
740 para_0<-c(alpha_0,beta_0,theta_0,family=family),delta_0) # Parámetros iniciales
741 a<-now()
742 para_optim <- optim(para_0,f_aux_reg,method ="BFGS",hessian = TRUE) # Estimación Maxima verosímil de los
743     parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
744 b<-now()
745 b-a #3.485547 hours
746 #load("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
747     Parcial/JointModel_Gaus.RData")
748 para_optim.value # Log_Likelihood
749 vector <- para_optim.par
750 p      <- ncol(var_y)
751 q      <- ncol(var_y)
752 Log_Likelihood<-para_optim.value
753 # Parámetros optimizados
754 alpha_gum<-vector[1:p]
755 beta_gum<-vector[(p+1):(p+q)]
756 theta_gum=z2theta(vector[p+q+1],family) #-0.09506928
757 delta_gum<-exp(vector[p+q+2]) #4.271691
758 tau_gum<-BiCopPar2Tau(par=theta_gum,family=family)#-0.06061453
759 head(beta_0)
760 head(beta_gum)
761 head(alpha_0)
762 head(alpha_gum)
763 # Determinamos los errores standar para el test de Wald
764 hessian_gum <- para_optim.hessian
765 Hinv       <- ginv(hessian_gum)
766 sd         <- sqrt(diag(Hinv))
767 sd.alpha_gum <-sd[1:p]
768 sd.beta_gum <- sd[(p+1):(p+q)]
769 sd.theta_gum <-sd[p+q+1]
770 # Ordenamos los coeficientes estimados en una tabla
771 Reg_Cop_Gumbel <- data.frame(Estimate_Frec=round(beta_gum,4),Std.error.freq=round(sd.beta_gum,4),z_value_
772     freq=round(beta_gum/sd.beta_gum,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_gum/sd.beta_gum))),4),Estimate_
773     Sev=round(alpha_gum,4),Std.error.sev=round(sd.alpha_gum,4),z_value_sev=round(alpha_gum/sd.alpha_gum,4),"_
774     Prob_Sev"=round(2*(1-pnorm(abs(alpha_gum/sd.alpha_gum))),4))
775 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
776 # Agregamos los interceptos
777 temp <- tibble(
778     var_level = c(
779         "Año2022",
780         "PT_Tipo_PolizaIndividual",
781         "PT_Uso_VehParticular",
782         "PT_Marca_VehL4",
783         "PT_ZonaA3",
784         "PT_GeneroFemenino",
785         "PT_Edad43-52"
786     ),
787     Estimate_Frec = 0
788     ,
789     Estimate_Sev = 0
790 )
791 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)_
792     `=round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))

```

```

794
795 Reg_Cop_Gumbel <-Reg_Cop_Gumbel %>% relocate(`exp(Estimate_Frec)`, .after=Prob_frec)
796
797 # Guardamos el modelo conjunto y los coeficientes
798 write.csv(Reg_Cop_Gumbel, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
799 GLM_Conjuntos/Perdida_Total/Coef_Gumb.csv", fileEncoding = "Latin1")
800 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
801 Conjuntos/Perdida_Total/JointModel_Gumb.RData")
802
803 # Simulacion de la Pérdida Total
804 # =====
805 # CON DEPENDENCIA
806 # -----
807 lambda_loss_dep <- exp(var_x%*%beta_clay)*data_CM.Expuesto
808 mu_loss_dep <- exp(var_x%*%alpha_clay)
809 delta_loss_dep <- delta_clay
810 # Función de densidad fx
811 f_loss <-
812   function(loss ,mu,d, lam ,theta){
813     n<-length(loss)
814     if (length(lam)==1) lam <- rep(lam,n)
815     if (length(mu)==1) mu <- rep(mu,n)
816     out <- vector(length=n)
817
818     for(i in 1:n){
819       N <-1:20
820       v <-rpois(N, lam[i])
821       vv <-rpois(N-1, lam[i])
822       u <-rgam(loss[i]/N,mu[i],d)
823
824       Der_cop <-Du_Clay(u,v, theta)-Du_Clay(u,vv, theta)
825       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N
826       out[i] <-sum(dummy)
827     }
828
829     out[loss <=0]=0
830     return(out)
831   }
832
833 # Función de densidad acumulada Fx
834 F_loss <-
835   function(loss ,mu,d, lam ,theta){
836     out<-vector(length=length(loss))
837
838     for(i in 1:length(loss)){
839       floss <- function(s){
840         f_loss(s,mu[i],d, lam[i], theta)
841       }
842       out[i] <- integrate(floss ,0 ,loss[i]).value
843     }
844     return(out)
845   }
846
847 # Estimación del Total Loss (fuerte dedicacion computacional)
848
849 #k <-length(lambda_loss_dep)
850 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
851 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
852 L <- vector(length=k)
853 S <- vector(length=m)
854
855 for(j in 1:m){
856   r_uni<-runif(k)
857   for(i in 1:k){
858     f_root <-function(s){
859       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_clay)-r_uni[i]
860     }
861
862     tryCatch(
863       error = function(cond) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
864       loss<-uniroot(f_root ,lower = 0,upper = 500000).root
865     )
866
867   #print(loss.root)

```

```

868     L[i]<-loss
869     perc<-paste(i/k*100,"%")
870     print(perc)
871   }
872   S[j]<-sum(L)
873   perc<-paste(j/m*100,"%")
874   print(perc)
875   #print(S[j])
876 }
877 S
878
879 # CON INDEPENDENCIA
880 # -----
881 lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
882 RIESGO COLECTIVO
883 mu_loss      <-predict(glm.Sev.PP, newdata=data_CM, type='response')
884 delta_loss    <-delta_0
885
886 M<-seq(1:10000)
887 for(i in 1:10000){
888   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
889   S<-vector(length=length(lambda_loss))
890   for(j in 1:length(muestra_cant)){
891     if(muestra_cant[j]==0){
892       S[j]=0
893     } else{
894       S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss))
895     }
896   }
897   M[i] <- sum(S)
898   porc<-paste((i/10000)*100,"%")
899   print(porc)
900   #print(M[i])
901 }
902
903 hist(M)
904
905 # Simulacion de la Frecuencia con Independencia
906 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
907 # Frecuencia
908 N<-seq(1:10000)
909 for(i in 1:10000){
910   N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
911   porc<-paste((i/10000)*100,"%")
912   print(porc)
913 }
914 hist(N)
915
916
917 ######
918 # Limpiamos el ambiente de variables
919 rm(list = ls())
920 gc()
921
922 #####
923 # REGRESION FRANK
924 #####
925
926 # Set Directory
927 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Perdida_"
928 Total")
929
930 # Carga los modelos GLM MARGINALES calculados en otros querys
931 load("glm.Frec.PT.RData")
932 load("glm.Sev.PT.RData")
933 load("Data_PT.RData")
934 load("Data_PT_model.RData")
935 load("Data_PT_model_CM.RData")
936
937 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
938 # SEVERIDAD
939 x <- data_CM.Incurrido/data_CM.Cantidad #(Costo Medio)
940 var_x <-model.matrix(glm.Sev.PT) #data_CM[,3:12]
941 # FRECUENCIA
942 y <- data_model.Cantidad

```

```

942 Expuestos <- data_model.Expuesto
943 var_y <- model.matrix(glm.Freq.PT) #data[,3:12]
944
945 # Estimacion de parámetros iniciales utilizando las distribuciones marginales
946 # SEVERIDAD
947 sd_alpha <- sqrt(diag(vcov(glm.Sev.PT)))
948 alpha_0 <- glm.Sev.PT.coefficients
949 delta_0 <- summary(glm.Sev.PT).dispersion
950 mu_0 <- exp(var_x%*%alpha_0)
951
952 # FRECUENCIA
953 beta_0 <- glm.Freq.PT.coefficients
954 sd_beta <- sqrt(diag(vcov(glm.Freq.PT)))
955 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
956
957 # REGRESION MEDIANTE COPULAS
958
959 # COPULA GUMBEL
960 family=5
961 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MLE
962 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
lambda_0)/(length(y)+1), family=family).par
963 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
964 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
965
966 # Estimación de las seudo-observaciones
967 u<-pgam(x,mu_0,delta_0)
968 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
969 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
970
971 # Creamos la función primera derivada de la Función Copula
972 Du_Frank <- function(u,v,theta){
973   u[u<=0]=0.0001
974   u[u>1]=1
975   v[v<=0]=0.0001
976   v[v>1]=1
977   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
978   #out <- (u^(1-theta)+v^(1-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
979   #out <- (u^(1-theta)+v^(1-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
980   out <- (exp(theta)*(exp(theta*v)-1))/(exp(theta*(u+1))+exp(theta*(v+1))-exp(theta)-exp(theta*(u+v)))
981   return(out)
982 }
983
984 # Función a maximizar
985 f_aux <- function(para){
986   D_u_0<-Du_Frank(u,v,para)-Du_Frank(u,vv,para)
987   D_u_0[D_u_0<=0]=1
988   D_u_0 <- log(D_u_0)
989   out<-(-sum(D_u_0))
990   return(out)
991 }
992
993 # Proceso de maximización
994 #para_ini <- theta2z(theta_ini,family = family)
995 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
996 #theta_0 <- z2theta(para_ifm,family = family)
997 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
# para la regresión conjunta
998 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
999
1000 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
1001
1002 # Creamos nuestra función a maximizar
1003 ##########
1004 f_aux_reg <- function(para){
1005
1006   p <- ncol(var_x)
1007   q <- ncol(var_x)
1008   alpha <- para[1:p]
1009   beta <- para[(p+1):(p+q)]
1010   # theta <- z2theta(para[p+q+1],family = family)
1011   theta <- z2theta(para[p+q+1],family = family)
1012   delta <- para[p+q+2]
1013   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
1014   mu <- as.vector(exp(var_y%*%alpha))

```

```

1015      #dummy    <- density_joint(x,y,mu,delta ,lambda ,theta ,family=family , zt=FALSE)
1016      u2 <- pgam(data_model.Incurrido ,mu, delta )
1017      v2<-ppois(y, lambda )
1018      vv2<-ppois(y-1,lambda )
1019
1020      marginal.x <- dgam(data_model.Incurrido ,mu, delta )
1021      marginal.x[marginal.x>=1]=0
1022      marginal.x[marginal.x<=0]=0
1023      marginal.x[is.na(marginal.x)]=0
1024      marginal.x[marginal.x==Inf]=0
1025      marginal.x[marginal.x==Inf]=0
1026
1027      par_der    <- Du_Frank(u2,v2,theta )
1028      par_der1   <- Du_Frank(u2,vv2,theta )
1029
1030      dummy<- par_der-par_der1
1031
1032      dummy[y==0]=par_der[y==0]
1033
1034      out<-marginal.x*dummy
1035
1036      out[out<=0]=1e-10
1037      ll       <- -sum(log(out))
1038
1039      #if(negative==TRUE) ll <- (-ll)
1040      return(ll)
1041  }
1042
1043 #####(#####
1044 para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
1045 a<-now()
1046 para_optim<- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Estimación Maxima verosmil de los
1047     parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
1048 b<-now()
1049 b-a #2.82135 hours
#load("C:/Users/josephgarcia1/OneDrive – KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
1050     Parcial/JointModel_Gaus.RData")
para_optim.value # Log_Likelihood
1051 vector <- para_optim.par
1052 p       <- ncol(var_y)
1053 q       <- ncol(var_y)
1054 Log_Likelihood<-para_optim.value
1055 # Parámetros optimizados
1056 alpha_frank<-vector[1:p]
1057 beta_frank<-vector[(p+1):(p+q)]
1058 theta_frank=z2theta(vector[p+q+1],family)
1059 delta_frank<-exp(vector[p+q+2]) #4.271691
1060 tau_frank<-BiCopPar2Tau(para=theta_frank,family=family)
1061 head(beta_0)
1062 head(beta_frank)
1063 head(alpha_0)
1064 head(alpha_frank)
1065
1066 # Determinamos los errores estandar para el test de Wald
1067 hessian_frank <- para_optim.hessian
1068 Hinv        <- ginv(hessian_frank)
1069 sd          <- sqrt(diag(Hinv))
1070 sd.alpha_frank <-sd[1:p]
1071 sd.beta_frank <- sd[(p+1):(p+q)]
1072 sd.theta_frank <-sd[p+q+1]
1073
1074 # Ordenamos los coeficientes estimados en una tabla
1075 Reg_Cop_Frank <- data.frame(Estimate_Frec=round(beta_frank,4),Std.error.freq=round(sd.beta_frank,4),z_
1076     value_freq=round(beta_frank/sd.beta_frank,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_frank/sd.beta_frank))),4),Estimate_Sev=round(alpha_frank,4),Std.error.sev=round(sd.alpha_frank,4),z_value_sev=round(alpha_frank/
1077     /sd.alpha_frank,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_frank/sd.alpha_frank))),4))
1078
1079 var_level<-row.names(Reg_Cop_Frank)
1080 Reg_Cop_Frank <- Reg_Cop_Frank %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
1081
1082 # Agregamos los interceptos
1083 temp <- tibble(
1084     var_level = c(
1085         "Año2022",
1086         "PT_Tipo_PolizaIndividual",
1087         "PT_Uso_Vehicular"
1088     )
1089 )

```

```

1086      "PT_Marca_VehL4",
1087      "PT_ZonaA3",
1088      "PT_GeneroFemenino",
1089      "PT_Edad43-52"
1090    ),
1091    Estimate_Frec = 0
1092  ,
1093  Estimate_Sev = 0
1094 )
1095
1096 # coeficientes del modelo de regresion con dependencia
1097 Reg_Cop_Frank <- Reg_Cop_Frank %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)`=
1098   round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
1099 Reg_Cop_Frank <-Reg_Cop_Frank %>% relocate(`exp(Estimate_Frec)` , .after=Prob_frec)
1100
1101 # Guardamos el modelo conjunto y los coeficientes
1102 write.csv(Reg_Cop_Frank, file="C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
1103 GLM_Conjuntos/Perdida_Total/Coef_Frank.csv", fileEncoding = "Latin1")
1103 save(para_optim, file="C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
1104 Conjuntos/Perdida_Total/JointModel_Frank.RData")
1105
1106 # Simulacion de la Pérdida Total
1107 # =====
1108 # CON DEPENDENCIA
1109 #
1110 lambda_loss_dep <- exp(var_x%*%beta_frank)*data_CM.Expuesto
1111 mu_loss_dep <- exp(var_x%*%alpha_frank)
1112 delta_loss_dep <- delta_frank
1113 # Función de densidad fx
1114 f_loss <-
1115   function(loss ,mu,d, lam ,theta){
1116     n<-length(loss)
1117     if (length(lam)==1) lam <- rep(lam ,n)
1118     if (length(mu)==1) mu <- rep(mu,n)
1119     out <- vector(length=n)
1120
1121     for(i in 1:n){
1122       N <-1:20
1123       v <-ppois(N, lam[i])
1124       vv <-ppois(N-1, lam[i])
1125       u <-pgam(loss[i]/N,mu[i] ,d)
1126
1127       Der_cop <-Du_Frank(u,v,theta)-Du_Frank(u,vv,theta) # Primera derivada de la función copula
1128       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N # Teorela del Total Loss
1129       out[i] <-sum(dummy)
1130     }
1131
1132     out[loss <=0]=0
1133     return(out)
1134   }
1135
1136 # Función de densidad acumulada Fx
1137 F_loss <-
1138   function(loss ,mu,d, lam ,theta){
1139     out<-vector(length=length(loss))
1140
1141     for(i in 1:length(loss)){
1142       floss <- function(s){
1143         f_loss(s,mu[i],d, lam[i],theta)
1144       }
1145       out[i] <- integrate(floss ,0,loss[i]).value # Integral para el cálculo del a F(x) a partir de f(x)
1146     }
1147     return(out)
1148   }
1149
1150 # Estimación del Total Loss (fuerte dedicacion computacional)
1151
1152 #k <-length(lambda_loss_dep)
1153 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
1154 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
1155 L <- vector(length=k)
1156 S <- vector(length=m)
1157
1158 for(j in 1:m){

```

```

1159     r_uni<-runif(k)
1160     for(i in 1:k){
1161       f_root <-function(s){
1162         F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_frank)-r_uni[i]
1163       }
1164
1165       tryCatch(
1166         error = function(cond) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
1167         loss<-uniroot(f_root,lower = 0,upper = 50000).root
1168       )
1169
1170       #print(loss.root)
1171       L[i]<-loss
1172       perc<-paste(i/k*100,"%")
1173       print(perc)
1174     }
1175     S[j]<-sum(L)
1176     perc<-paste(j/m*100,"%")
1177     print(perc)
1178     #print(S[j])
1179   }
1180   S
1181
1182   # CON INDEPENDENCIA
1183   #
1184   lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
1185   # RIESGO COLECTIVO
1186   mu_loss <-predict(glm.Sev.PP, newdata=data_CM, type='response')
1187   delta_loss <-delta_0
1188
1189   M<-seq(1:10000)
1190   for(i in 1:10000){
1191     muestra_cant<-rpois(length(lambda_loss),lambda_loss)
1192     S<-vector(length=length(lambda_loss))
1193     for(j in 1:length(muestra_cant)){
1194       if(muestra_cant[j]==0){
1195         S[j]=0
1196       } else{
1197         S[j]=sum(rgam(n=muestra_cant[j],mu_loss[j],delta_loss))
1198       }
1199     }
1200     M[i] <- sum(S)
1201     porc<-paste((i/10000)*100,"%")
1202     print(porc)
1203     #print(M[i])
1204   }
1205
1206   hist(M)
1207   # Simulacion de la Frecuencia con Independencia
1208   #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
1209   # Frecuencia
1210   N<-seq(1:10000)
1211   for(i in 1:10000){
1212     N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
1213     porc<-paste((i/10000)*100,"%")
1214     print(porc)
1215   }
1216   hist(N)

```

C.5. Modelo Conjunto - Responsabilidad Civil

```

1 ##########
2 #####
3 #####
4 #####      REGRESION BASADO EN COPULAS      #####
5 #####      RESPONSABILIDAD CIVIL      #####
6 #####
7 #####
8 #####
9 #####
10 # Cargamos las librerías necesarias
11
12 #library(copula)
13 library(RODBC)

```

```

14 library(tidyverse)
15 library(MASS)
16 #library(GJRM)
17 #library(devtools)
18 library(CopulaRegression)
19 library(VineCopula)
20 #library(optimx)
21 #install_url('http://cran.r-project.org/src/contrib/Archive/CopulaRegression/CopulaRegression_0.1-5.tar.gz')
22 #####
23 # REGRESION GAUSSIANA
24 #####
25 #####
26 #####
27 # Set Directory
28 setwd("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
    Responsabilidad_Civil")
29 #####
30 # Carga los modelos GLM MARGINALES calculados en otros querys
31 load("glm.Frec.RData")
32 load("glm.Sev.RData")
33 load("Data_RC.RData")
34 load("Data_RC_model.RData")
35 load("Data_RC_model_CM.RData")
36 #####
37 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
38 # SEVERIDAD
39 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
40 var_x <- model.matrix(glm.Sev.RC) #data_CM[,3:12]
41 # FRECUENCIA
42 y <- data_model.Cantidad
43 Expuestos <- data_model.Expuesto
44 var_y <- model.matrix(glm.Frec.RC) #data[,3:12]
45 #####
46 # Estimacion de parametros iniciales utilizando las distribuciones marginales
47 # SEVERIDAD
48 sd_alpha <- sqrt(diag(vcov(glm.Sev.RC)))
49 alpha_0 <- glm.Sev.RC.coefficients
50 delta_0 <- summary(glm.Sev.RC).dispersion
51 mu_0 <- exp(var_x%*%alpha_0)
52 #####
53 # FRECUENCIA
54 beta_0 <- glm.Frec.RC.coefficients
55 sd_beta <- sqrt(diag(vcov(glm.Frec.RC)))
56 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
57 #####
58 # REGRESION MEDIANTE COPULAS
59 #####
60 # COPULA GAUSSIANA
61 family=1
62 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
63 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-lambda_0)/(length(y)+1), family=family).par
64 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
65 tau_ini <- BiCopPar2Tau(par=theta_ini, family = family)
66 #####
67 # Estimación de las seudo-observaciones
68 u<-pgam(x,mu_0,delta_0)
69 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
70 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
71 #####
72 # Creamos la función primera derivada de la Función Copula
73 Du_Gaus <- function(u,v,theta){
74   u[u<=0]=0.001
75   u[u>=1]=0.999
76   v[v<=0]=0.001
77   v[v>=1]=0.999
78   out <- pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
79   return(out)
80 }
81 #####
82 # Función a maximizar
83 f_aux <- function(para){
84   D_u_0<-Du_Gaus(u,v,para)-Du_Gaus(u,vv,para)
85   D_u_0[D_u_0<=0]=1
86   D_u_0 <- log(D_u_0)

```

```

87     out<-(-sum(D_u_0))
88     return(out)
89 }
90
91 # Proceso de maximización
92 #para_ini <- theta2z(theta_ini,family = family)
93 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
94 #theta_0 <- z2theta(para_ifm,family = family)
95 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
# para la regresión conjunta
96 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
97
98 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
99
100 # Creamos nuestra función a maximizar
101 ##########
102 f_aux_reg <- function(para){
103
104     p <- ncol(var_x)
105     q <- ncol(var_x)
106     alpha <- para[1:p]
107     beta <- para[(p+1):(p+q)]
108     # theta <- z2theta(para[p+q+1],family)
109     theta <- para[p+q+1]
110     delta <- para[p+q+2]
111     lambda <- as.vector(exp(var_y%*%beta)*data_model.Exposto)
112     mu <- as.vector(exp(var_y%*%alpha))
113     #dummy <- density_joint(x,y,mu,delta,lambda,theta,family=family,zt=FALSE)
114     u2 <- pgam(data_model.Incurrido,mu,delta)
115     v2<-ppois(y,lambda)
116     vv2<-ppois(y-1,lambda)
117
118     marginal.x <- dgam(data_model.Incurrido ,mu,delta)
119     marginal.x[marginal.x>=1]=0
120     marginal.x[marginal.x<=0]=0
121     marginal.x[is.na(marginal.x)]=0
122     marginal.x[marginal.x== -Inf]=0
123     marginal.x[marginal.x== Inf]=0
124
125     par_der <- Du_Gaus(u2,v2 ,theta )
126     par_der1 <- Du_Gaus(u2,vv2 ,theta )
127
128     dummy<- par_der-par_der1
129
130     dummy[y==0]=par_der[y==0]
131
132     out<-marginal.x*dummy
133
134     out[out<=0]=1e-10
135     ll <- -sum(log(out))
136
137     #if(negative==TRUE) ll <- (-ll)
138     return(ll)
139 }
140
141 #########
142
143 para_0<-c(alpha_0,beta_0,theta_0,delta_0) # Parámetros iniciales
144 a<-now()
145 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Demora 6min
146 b<-now()
147 b-a
148 #load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/JointModel_Gaus.RData")
149 para_optim$value # Log_Likelihood
150 vector <- para_optim$par
151 p <- ncol(var_y)
152 q <- ncol(var_y)
153
154 # Parámetros optimizados
155 alpha_gau<-vector[1:p]
156 beta_gau<-vector[(p+1):(p+q)]
157 theta_gau=z2theta(vector[p+q+1],family)
158 delta_gau<-exp(vector[p+q+2])
159 tau_gau<-BiCopPar2Tau(par=theta_gau,family=family)
160 head(beta_0)

```

```

161 head(beta_gau)
162 head(alpha_0)
163 head(alpha_gau)
164 # Determinamos los errores estandar para el test de Wald
165 hessian_gau <- para_optim.hessian
166 Hinv <- ginv(hessian_gau)
167 sd <- sqrt(diag(Hinv))
168 sd.alpha_gau <-sd[1:p]
169 sd.beta_gau <- sd[(p+1):(p+q)]
170 sd.theta_gau <-sd[p+q+1]
171
172 # Ordenamos los coeficientes estimados en una tabla
173 Reg_Cop_Gausiana <- data.frame(Estimate_Frec=round(beta_gau,4),Std.error.freq=round(sd.beta_gau,4),z_value_frec=round(beta_gau/sd.beta_gau,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_gau/sd.beta_gau))),4),
Estimate_Sev=round(alpha_gau,4),Std.error.sev=round(sd.alpha_gau,4),z_value_sev=round(alpha_gau/sd.alpha_gau,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_gau/sd.alpha_gau))),4))
174 #Reg_Cop_Gausiana <- data.frame(Estimate_Frec=round(beta_gau,4),Std.error.freq=0,z_value_frec=0,"Prob_frec"=0,Estimate_Sev=round(alpha_gau,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
175 var_level<-row.names(Reg_Cop_Gausiana)
176 Reg_Cop_Gausiana <- Reg_Cop_Gausiana %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
177
178 # Agregamos los interceptos
179 temp <- tibble(
180   var_level = c(
181     "Año2022",
182     "RC_AntiguedadB_2-9",
183     "RC_Canal_VentaCorredores",
184     "RC_Uso_Vehicular",
185     "RC_Marca_VehL3",
186     "RC_ZonaA2",
187     "RC_Edad43-52"
188   ),
189   Estimate_Frec = 0
190   ,
191   Estimate_Sev = 0
192 )
193
194 # coeficientes del modelo de regresion con dependencia
195 Reg_Cop_Gausiana <- Reg_Cop_Gausiana %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)`=round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
196
197 Reg_Cop_Gausiana <-Reg_Cop_Gausiana %>% relocate(`exp(Estimate_Frec)`, .after=Prob_frec)
198
199 # Guardamos el modelo conjunto y los coeficientes
200 write.csv(Reg_Cop_Gausiana, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Responsabilidad_Civil/Coef_Gaus.csv", fileEncoding = "Latin1")
201 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Responsabilidad_Civil/JointModel_Gaus.RData")
202
203 # Simulacion de la Pérdida Total
204 # =====
205
206 # CON DEPENDENCIA
207 #
208 lambda_loss_dep <- exp(var_x%&%beta_gau)*data_CM.Expuesto
209 mu_loss_dep <- exp(var_x%&%alpha_gau)
210 delta_loss_dep <- delta_gau
211
212
213 # Función de densidad fx
214
215 f_loss <-
216   function(loss, mu, d, lam, theta){
217     n<-length(loss)
218     if (length(lam)==1) lam <- rep(lam,n)
219     if (length(mu)==1) mu <- rep(mu,n)
220     out <- vector(length=n)
221
222     for(i in 1:n){
223       N <-1:20
224       v <-ppois(N, lam[i])
225       vv <-ppois(N-1, lam[i])
226       u <-pgam(loss[i]/N, mu[i], d)
227
228       Der_cop <-Du_Gaus(u,v,theta)-Du_Gaus(u,vv,theta)

```

```

229      dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N
230      out[i] <-sum(dummy)
231    }
232
233    out[loss<=0]=0
234    return(out)
235  }
236
237 # Función de densidad acumulada Fx
238 F_loss <-
239   function(loss, mu, d, lam, theta){
240     out<-vector(length=length(loss))
241
242     for(i in 1:length(loss)){
243       floss <- function(s){
244         f_loss(s,mu[i],d, lam[i], theta)
245       }
246       out[i] <- integrate(floss, 0, loss[i]).value
247     }
248     return(out)
249   }
250
251 # Estimación del Total Loss (fuerte dedicacion computacional)
252
253 #k <-length(lambda_loss_dep)
254 k <- length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
255 m <- 100 # Cantidad de muestras aleatorias (Deben ser 10 mil)
256 L <- vector(length=k)
257 S <- vector(length=m)
258
259 for(j in 1:m){
260   r_uni<-runif(k)
261   for(i in 1:k){
262     f_root <-function(s){
263       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_gau)-r_uni[i]
264     }
265
266     tryCatch(
267       error = function(cond) mu_loss_dep[i]*lambda_loss_dep[i],
268       loss<-uniroot(f_root, lower = 0,upper = 500000)
269     )
270
271     #print(loss.root)
272     L[i]<-loss.root
273     perc<-paste(i/k*100,"%")
274     print(perc)
275   }
276   S[j]<-sum(L)
277   perc<-paste(j/m*100,"%")
278   print(perc)
279   #print(S[j])
280 }
281 head(S)
282
283 # CON INDEPENDENCIA
284 #
285 lambda_loss <-predict(glm.Frec.PT, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN
286 MODELO DE RIESGO COLECTIVO
287 mu_loss <-predict(glm.Sev.PT, newdata=data_CM, type='response')
288 delta_loss <-delta_0
289
290 M<-seq(1:10000)
291 for(i in 1:10000){
292   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
293   S<-vector(length=length(lambda_loss))
294   for(j in 1:length(muestra_cant)){
295     if(muestra_cant[j]==0){
296       S[j]=0
297     } else {
298       S[j]=sum(rgam(n=muestra_cant[j],mu_loss[j],delta_loss))
299     }
300
301   M[i] <- sum(S)
302   porc<-paste((i/10000)*100,"%")
303   print(porc)

```

```

304         # print(M[i])
305     }
306
307     hist(M, breaks = 50)
308     mean(M)
309     sum(data.Incurrido)
310
311
312 # Simulacion de la Frecuencia con Independencia
313 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
314 # Frecuencia
315 N<-seq(1:10000)
316 for(i in 1:10000){
317   N[i] <- sum(simulate(glm.Freq.PT,nsim=1, type="response") .sim_1)/sum(data.model.Expuesto)
318   porc<-paste((i/10000)*100,"%")
319   print(porc)
320 }
321 hist(N, breaks = 50)
322 mean(N)
323 mean(data.Frecuencia)
324
325 #####
326 # Limpiamos el ambiente de variables
327 rm(list = ls())
328 gc()
329
330 #####
331 # REGRESION CLAYTON
332 #####
333
334 # Set Directory
335 setwd("C:/Users/josephgarcia /OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
      Responsabilidad_Civil")
336
337 # Carga los modelos GLM MARGINALES calculados en otros querys
338 load("glm.Freq.RData")
339 load("glm.Sev.RData")
340 load("Data.RData")
341 load("Data_RC_model.RData")
342 load("Data_RC_model_CM.RData")
343
344 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
345 # SEVERIDAD
346 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
347 var_x <- model.matrix(glm.Sev.RC) #data_CM[,3:12]
348 # FRECUENCIA
349 y <- data_model.Cantidad
350 Expuestos <- data_model.Expuesto
351 var_y <- model.matrix(glm.Freq.RC) #data[,3:12]
352
353 # Estimacion de parametros iniciales utilizando las distribuciones marginales
354 # SEVERIDAD
355 sd.alpha <- sqrt(diag(vcov(glm.Sev.RC)))
356 alpha_0 <- glm.Sev.RC.coefficients
357 delta_0 <- summary(glm.Sev.RC).dispersion
358 mu_0 <- exp(var_x%*%alpha_0)
359
360 # FRECUENCIA
361 beta_0 <- glm.Freq.RC.coefficients
362 sd.beta <- sqrt(diag(vcov(glm.Freq.RC)))
363 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
364
365 # REGRESION MEDIANTE COPULAS
366
367 # COPULA CLAYTON
368 family=3
369 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
370 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
    lambda_0)/(length(y)+1), family=family).par
371 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*
    data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
372 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
373
374 # Estimación de las seudo-observaciones
375 u<-pgam(x,mu_0,delta_0)
376 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)

```

```

377 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%%beta_0)*data_CM.Expuesto)
378
379 # Creamos la función primera derivada de la Función Copula
380 Du_Clay <- function(u,v,theta){
381   u[u<=0]=0.0001
382   u[u>1]=1
383   v[v<=0]=0.0001
384   v[v>1]=1
385   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
386   out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta)-1))
387   return(out)
388 }
389
390 # Función a maximizar
391 f_aux <- function(para){
392   D_u_0<-Du_Clay(u,v,para)-Du_Clay(u,vv,para)
393   D_u_0[D_u_0<=0]=1
394   D_u_0 <- log(D_u_0)
395   out<-(-sum(D_u_0))
396   return(out)
397 }
398
399 # Proceso de maximización
400 #para_ini <- theta2z(theta_ini,family = family)
401 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
402 #theta_0 <- z2theta(para_ifm,family = family)
403 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
404 # para la regresión conjunta
405 tau <- BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
406
407 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
408
409 # Creamos nuestra función a maximizar
410 #####f_aux_reg <- function(para){
411
412   p <- ncol(var_x)
413   q <- ncol(var_x)
414   alpha <- para[1:p]
415   beta <- para[(p+1):(p+q)]
416   # theta <- z2theta(para[p+q+1],family)
417   theta <- z2theta(para[p+q+1],family = family)
418   delta <- para[p+q+2]
419   lambda <- as.vector(exp(var_y%%beta)*data_model.Expuesto)
420   mu <- as.vector(exp(var_y%%alpha))
421   #dummy <- density_joint(x,y,mu,delta,lambda,theta,family=family,zt=FALSE)
422   u2 <- pgam(data_model.Incurrido,mu,delta)
423   v2<-ppois(y,lambda)
424   vv2<-ppois(y-1,lambda)
425
426   marginal.x <- dgam(data_model.Incurrido,mu,delta)
427   marginal.x[marginal.x>=1]=0
428   marginal.x[marginal.x<=0]=0
429   marginal.x[is.na(marginal.x)]=0
430   marginal.x[marginal.x== -Inf]=0
431   marginal.x[marginal.x== Inf]=0
432
433   par_der <- Du_Clay(u2,v2,theta)
434   par_der1 <- Du_Clay(u2,vv2,theta)
435
436   dummy<- par_der-par_der1
437
438   dummy[y==0]=par_der[y==0]
439
440   out<-marginal.x*dummy
441
442   out[out<=0]=1e-10
443   ll <- -sum(log(out))
444
445   #if(negative==TRUE) ll <- (-ll)
446   return(ll)
447 }
448
449 #####para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
450 a<-now()
451

```

```

452 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Estimación Maxima verosmil de los
453                                         # parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
454 b<-now()
455 b-a
456 #load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
457                                         Parcial/JointModel_Gaus.RData")
458 para_optim.value # Log_Likelihood
459 vector <- para_optim.par
460 p               <- ncol(var_y)
461 q               <- ncol(var_y)
462 Log_Likelihood<-para_optim.value
463 # Parámetros optimizados
464 alpha_clay<-vector[1:p]
465 beta_clay<-vector[(p+1):(p+q)]
466 theta_clay=z2theta(vector[p+q+1],family) #-0.09506928
467 delta_clay<-exp(vector[p+q+2]) #4.271691
468 tau_clay<-BiCopPar2Tau(par=theta_clay,family=family)#-0.06061453
469 head(beta_0)
470 head(beta_clay)
471 head(alpha_0)
472 head(alpha_clay)
473 # Determinamos los errores standar para el test de Wald
474 hessian_clay <- para_optim.hessian
475 Hinv          <- ginv(hessian_clay)
476 sd             <- sqrt(diag(Hinv))
477 sd.alpha_clay <-sd[1:p]
478 sd.beta_clay  <- sd[(p+1):(p+q)]
479 sd.theta_clay <-sd[p+q+1]
480
481 # Ordenamos los coeficientes estimados en una tabla
482 Reg_Cop_Clayton <- data.frame(Estimate_Frec=round(beta_clay,4),Std.error.freq=round(sd.beta_clay,4),z_
483                                         value_frec=round(beta_clay/sd.beta_clay,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_clay/sd.beta_clay))),4),
484                                         Estimate_Sev=round(alpha_clay,4),Std.error.sev=round(sd.alpha_clay,4),z_value_sev=round(alpha_clay/sd.
485                                         alpha_clay,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_clay/sd.alpha_clay))),4))
486 #Reg_Cop_Clayton <- data.frame(Estimate_Frec=round(beta_clay,4),Std.error.freq=0,z_value_frec=0,"Prob_frec"
487                                         ="0",Estimate_Sev=round(alpha_clay,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
488 var_level<-row.names(Reg_Cop_Clayton)
489 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level
490                                         )
491
492 # Agregamos los interceptos
493 temp <- tibble(
494   var_level = c(
495     "Año2022",
496     "RC_Antiguedad_b_2-9",
497     "RC_Canal_VentaCorredores",
498     "RC_Uso_VehParticular",
499     "RC_Marca_VehL3",
500     "RC_ZonaA2",
501     "RC_Edad43-52"
502   ),
503   Estimate_Frec = 0
504   ,
505   Estimate_Sev = 0
506 )
507
508 # coeficientes del modelo de regresion con dependencia
509 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_
510                                         Frec)`=round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
511
512 Reg_Cop_Clayton <-Reg_Cop_Clayton %>% relocate(`exp(Estimate_Frec)`,.after=Prob_frec)
513
514 # Guardamos el modelo conjunto y los coeficientes
515 write.csv(Reg_Cop_Clayton, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
516                                         GLM_Conjuntos/Responsabilidad_Civil/Coef_Clay.csv", fileEncoding = "Latin1")
517 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
518                                         Conjuntos/Responsabilidad_Civil/JointModel_Clay.RData")
519
520 # Simulacion de la Pérdida Total
521 # =====
522
523 # CON DEPENDENCIA
524 # -----
525 lambda_loss_dep <- exp(var_x%*%beta_clay)*data_CM.Expuesto
526 mu_loss_dep    <- exp(var_x%*%alpha_clay)
527 delta_loss_dep <- delta_clay

```

```

518
519 # Función de densidad fx
520
521 f_loss <-
522   function(loss ,mu,d, lam ,theta){
523     n<-length(loss)
524     if (length(lam)==1) lam <- rep(lam ,n)
525     if (length(mu)==1) mu <- rep(mu,n)
526     out <- vector(length=n)
527
528     for(i in 1:n){
529       N <-1:20
530       v <-ppois(N, lam [ i ])
531       vv <-ppois(N-1, lam [ i ])
532       u <-pgam(loss [ i ]/N,mu[ i ],d)
533
534       Der_cop <-Du_Clay(u,v, theta)-Du_Clay(u,vv, theta)
535       dummy <-Der_cop*dgam( loss [ i ]/N,mu[ i ],d)/N
536       out [ i ] <-sum(dummy)
537     }
538
539     out [ loss <=0]=0
540     return(out)
541   }
542
543
544 # Función de densidad acumulada Fx
545 F_loss <-
546   function(loss ,mu,d, lam ,theta){
547     out<-vector(length=length(loss))
548
549     for(i in 1:length(loss)){
550       floss <- function(s){
551         f_loss(s,mu[ i ],d, lam [ i ],theta)
552       }
553       out [ i ] <- integrate(floss ,0 ,loss [ i ]). value
554     }
555     return(out)
556   }
557
558 # Estimación del Total Loss (fuerte dedicacion computacional)
559
560 #k <-length(lambda_loss_dep)
561 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
562 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
563 L <- vector(length=k)
564 S <- vector(length=m)
565
566 for(j in 1:m){
567   r_uni<-runif(k)
568   for(i in 1:k){
569     f_root <-function(s){
570       F_loss(s,mu_loss_dep [ i ],delta_loss_dep ,lambda_loss_dep [ i ],theta_clay)-r_uni [ i ]
571     }
572
573     tryCatch(
574       error = function(cnd) loss<-mu_loss_dep [ i ]*lambda_loss_dep [ i ],
575       loss<-uniroot(f_root ,lower = 0 ,upper = 500000). root
576     )
577
578     #print(loss .root)
579     L [ i ]<-loss
580     perc<-paste(i/k*100,"%")
581     print(perc)
582   }
583   S [ j ]<-sum(L)
584   perc<-paste(j/m*100,"%")
585   print(perc)
586   #print(S [ j ])
587 }
588 S
589
590 # CON INDEPENDENCIA
591 # -----
592 lambda_loss <-predict(glm.Frec_PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO

```

```

593 mu_loss      <-predict(glm.Sev.PP, newdata=data_CM, type='response')
594 delta_loss   <-delta_0
595
596 M<-seq(1:10000)
597 for(i in 1:10000){
598   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
599   S<-vector(length=length(lambda_loss))
600   for(j in 1:length(muestra_cant)){
601     if(muestra_cant[j]==0){
602       S[j]=0
603     } else{
604       S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
605     }
606   }
607
608   M[i] <- sum(S)
609   porc<-paste((i/10000)*100,"%")
610   print(porc)
611   #print(M[i])
612 }
613
614 hist(M)
615
616 # Simulacion de la Frecuencia con Independencia
617 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
618 # Frecuencia
619 N<-seq(1:10000)
620 for(i in 1:10000){
621   N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
622   porc<-paste((i/10000)*100,"%")
623   print(porc)
624 }
625 hist(N)
626 ######
627 # Limpiamos el ambiente de variables
628 rm(list = ls())
629 gc()
630
631 #####
632 # REGRESION GUMBEL
633 #####
634
635 # Set Directory
636 setwd("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
  Responsabilidad_Civil")
637
638 # Carga los modelos GLM MARGINALES calculados en otros querys
639 load("glm.Frec.RData")
640 load("glm.Sev.RData")
641 load("Data_RC.RData")
642 load("Data_RC_model.RData")
643 load("Data_RC_model_CM.RData")
644
645 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
646 # SEVERIDAD
647 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
648 var_x <- model.matrix(glm.Sev.RC) #data_CM[,3:12]
649 # FRECUENCIA
650 y <- data_model.Cantidad
651 Expuestos <- data_model.Expuesto
652 var_y <- model.matrix(glm.Frec.RC) #data[,3:12]
653
654 # Estimacion de parametros iniciales utilizando las distribuciones marginales
655 # SEVERIDAD
656 sd_alpha <- sqrt(diag(vcov(glm.Sev.RC)))
657 alpha_0 <- glm.Sev.RC.coefficients
658 delta_0 <- summary(glm.Sev.RC).dispersion
659 mu_0 <- exp(var_x%*%alpha_0)
660
661 # FRECUENCIA
662 beta_0 <- glm.Frec.RC.coefficients
663 sd_beta <- sqrt(diag(vcov(glm.Frec.RC)))
664 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
665
666 # REGRESION MEDIANTE COPULAS
667

```

```

668 # COPULA GUMBEL
669 family=4
670 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
671 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
  lambda_0)/(length(y)+1), family=family).par
672 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
  data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
673 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
674
675 # Estimación de las seudo-observaciones
676 u<-pgam(x,mu_0,delta_0)
677 v <- ppois(data_CM.Cantidad ,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
678 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
679
680 # Creamos la función primera derivada de la Función Copula
681 Du_Gum <- function(u,v,theta){
682   u[u<=0]=0.0001
683   u[u>1]=1
684   v[v<=0]=0.0001
685   v[v>1]=1
686   #out <- pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
687   #out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
688   out <- (u^-1)*exp(-((-log(u))^theta+(-log(v))^theta)^(1/theta))
689   return(out)
690 }
691
692 # Función a maximizar
693 f_aux <- function(para){
694   D_u_0<-Du_Gum(u,v,para)-Du_Gum(u,vv,para)
695   D_u_0[D_u_0<=0]=1
696   D_u_0 <- log(D_u_0)
697   out<-(sum(D_u_0))
698   return(out)
699 }
700
701 # Proceso de maximización
702 #para_ini <- theta2z(theta_ini,family = family)
703 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
704 #theta_0 <- z2theta(para_ifm,family = family)
705 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
  para la regresión conjunta
706 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
707
708 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
709
710 # Creamos nuestra función a maximizar
711 #####f_aux_reg <- function(para){
712
713   p <- ncol(var_x)
714   q <- ncol(var_x)
715   alpha <- para[1:p]
716   beta <- para[(p+1):(p+q)]
717   # theta <- z2theta(para[p+q+1],family)
718   theta <- z2theta(para[p+q+1],family = family)
719   delta <- para[p+q+2]
720   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
721   mu <- as.vector(exp(var_y%*%alpha))
722   #dummy <- density_joint(x,y,mu,delta ,lambda ,theta ,family=family ,zt=FALSE)
723   u2 <- pgam(data_model.Incurrido ,mu, delta )
724   v2<-ppois(y,lambda)
725   vv2<-ppois(y-1,lambda)
726
727   marginal.x <- dgam(data_model.Incurrido ,mu, delta )
728   marginal.x[marginal.x>=1]=0
729   marginal.x[marginal.x<=0]=0
730   marginal.x[is.na(marginal.x)]=0
731   marginal.x[marginal.x== -Inf]=0
732   marginal.x[marginal.x== Inf]=0
733
734   par_der <- Du_Gum(u2,v2,theta)
735   par_der1 <- Du_Gum(u2,vv2,theta)
736
737   dummy<- par_der-par_der1
738
739   dummy[y==0]=par_der[y==0]
740

```

```

741     out<-marginal.x*dummy
742
743     out[ out<=0]=1e-10
744     ll      <- -sum(log(out))
745
746     #if(negative==TRUE) ll <- (-ll)
747     return(ll)
748   }
749
750 #####
751 para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
752 a<-now()
753 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Estimación Maxima verosímil de los
754     parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
755 b<-now()
756 b-a #3.485547 hours
757 #load("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
758     _Parcial/JointModel_Gaus.RData")
759 para_optim.value # Log_Likelihood
760 vector <- para_optim.par
761 p       <- ncol(var_y)
762 q       <- ncol(var_y)
763 Log_Likelihood<-para_optim.value
764 # Parámetros optimizados
765 alpha_gum<-vector[1:p]
766 beta_gum<-vector[(p+1):(p+q)]
767 theta_gum=z2theta(vector[p+q+1],family)
768 delta_gum<-exp(vector[p+q+2])
769 tau_gum<-BiCopPar2Tau(par=theta_gum,family=family)
770 head(beta_0)
771 head(beta_gum)
772 head(alpha_0)
773 head(alpha_gum)
774 # Determinamos los errores standar para el test de Wald
775 hessian_gum <- para_optim.hessian
776 Hinv        <- ginv(hessian_gum)
777 sd          <- sqrt(diag(Hinv))
778 sd.alpha_gum <-sd[1:p]
779 sd.beta_gum <- sd[(p+1):(p+q)]
780 sd.theta_gum <-sd[p+q+1]
781
782 # Ordenamos los coeficientes estimados en una tabla
783 Reg_Cop_Gumbel <- data.frame(Estimate_Frec=round(beta_gum,4),Std.error.freq=round(sd.beta_gum,4),z_value_
784     _freq=round(beta_gum/sd.beta_gum,4),"Prob_freq"=round(2*(1-pnorm(abs(beta_gum/sd.beta_gum))),4),Estimate_
785     _Sev=round(alpha_gum,4),Std.error.sev=round(sd.alpha_gum,4),z_value_sev=round(alpha_gum/sd.alpha_gum,4),"_
786     Prob_Sev"=round(2*(1-pnorm(abs(alpha_gum/sd.alpha_gum))),4))
787 #Reg_Cop_Gumbel <- data.frame(Estimate_Frec=round(beta_gum,4),Std.error.freq=0,z_value_freq=0,"Prob_freq"
788     "=0,Estimate_Sev=round(alpha_gum,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
789 var_level<-row.names(Reg_Cop_Gumbel)
790 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
791
792 # Agregamos los interceptos
793
794 temp <- tibble(
795   var_level = c(
796     "Año2022",
797     "RC_AntiguedadB_2-9",
798     "RC_Canal_VentaCorredores",
799     "RC_Uso_VehParticular",
800     "RC_Marca_VehL3",
801     "RC_ZonaA2",
802     "RC_Edad43-52"
803   ),
804   Estimate_Frec = 0
805   ,
806   Estimate_Sev = 0
807 )
808
809 # coeficientes del modelo de regresion con dependencia
810 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)`
811     `=round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
812
813 Reg_Cop_Gumbel <-Reg_Cop_Gumbel %>% relocate(`exp(Estimate_Frec)`,.after=Prob_freq)

```

```

810 # Guardamos el modelo conjunto y los coeficientes
811 write.csv(Reg_Cop_Gumbel, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
812 GLM_Conjuntos/Responsabilidad_Civil/Coef_Gumb.csv", fileEncoding = "Latin1")
813 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
814 Conjuntos/Responsabilidad_Civil/JointModel_Gumb.RData")
815
816 # Simulacion de la Pérdida Total
817 # =====
818 # CON DEPENDENCIA
819 # -----
820 lambda_loss_dep <- exp(var_x%*%beta_clay)*data_CM.Expuesto
821 mu_loss_dep <- exp(var_x%*%alpha_clay)
822 delta_loss_dep <- delta_clay
823
824 # Función de densidad fx
825
826 f_loss <-
827   function(loss ,mu,d, lam ,theta){
828     n<-length(loss)
829     if (length(lam)==1) lam <- rep(lam,n)
830     if (length(mu)==1) mu <- rep(mu,n)
831     out <- vector(length=n)
832
833     for(i in 1:n){
834       N <-1:20
835       v <-ppois(N, lam[i])
836       vv <-ppois(N-1, lam[i])
837       u <-pgam(loss[i]/N,mu[i],d)
838
839       Der_cop <-Du_Clay(u,v, theta)-Du_Clay(u,vv, theta)
840       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N
841       out[i] <-sum(dummy)
842     }
843
844     out[loss <=0]=0
845     return(out)
846   }
847
848 # Función de densidad acumulada Fx
849 F_loss <-
850   function(loss ,mu,d, lam ,theta){
851     out<-vector(length=length(loss))
852
853     for(i in 1:length(loss)){
854       floss <- function(s){
855         f_loss(s,mu[i],d, lam[i], theta)
856       }
857       out[i] <- integrate(floss ,0 ,loss[i]).value
858     }
859     return(out)
860   }
861
862 # Estimación del Total Loss (fuerte dedicacion computacional)
863
864 #k <-length(lambda_loss_dep)
865 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
866 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
867 L <- vector(length=k)
868 S <- vector(length=m)
869
870 for(j in 1:m){
871   r_uni<-runif(k)
872   for(i in 1:k){
873     f_root <-function(s){
874       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_clay)-r_uni[i]
875     }
876
877     tryCatch(
878       error = function(cnd) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
879       loss<-uniroot(f_root ,lower = 0,upper = 500000).root
880     )
881
882     #print(loss.root)
883     L[i]<-loss

```

```

884     perc<-paste(i/k*100,"%")
885     print(perc)
886   }
887   S[j]<-sum(L)
888   perc<-paste(j/m*100,"%")
889   print(perc)
890   #print(S[j])
891 }
892 S
893
894 # CON INDEPENDENCIA
895 # -----
896 lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO
897 mu_loss      <-predict(glm.Sev.PP, newdata=data_CM, type='response')
898 delta_loss   <-delta_0
899
900 M<-seq(1:10000)
901 for(i in 1:10000){
902   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
903   S<-vector(length=length(lambda_loss))
904   for(j in 1:length(muestra_cant)){
905     if(muestra_cant[j]==0){
906       S[j]=0
907     } else {
908       S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
909     }
910   }
911
912   M[i] <- sum(S)
913   porc<-paste((i/10000)*100,"%")
914   print(porc)
915   #print(M[i])
916 }
917 hist(M)
918
919 # Simulacion de la Frecuencia con Independencia
920 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
921 # Frecuencia
922 N<-seq(1:10000)
923 for(i in 1:10000){
924   N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
925   porc<-paste((i/10000)*100,"%")
926   print(porc)
927 }
928 hist(N)
929 #####
930 # Limpiamos el ambiente de variables
931 rm(list = ls())
932 gc()
933
934 #####
935 #####
936 # REGRESION FRANK
937 #####
938
939 # Set Directory
940 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/
Responsabilidad_Civil")
941
942 # Carga los modelos GLM MARGINALES calculados en otros querys
943 load("glm.Frec.RData")
944 load("glm.Sev.RData")
945 load("Data_RC.RData")
946 load("Data_RC_model.RData")
947 load("Data_RC_model_CM.RData")
948
949 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
950 # SEVERIDAD
951 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
952 var_x <-model.matrix(glm.Sev.RC) #data_CM[,3:12]
953 # FRECUENCIA
954 y <- data_model.Cantidad
955 Expuestos <- data_model.Expuesto
956 var_y <- model.matrix(glm.Frec.RC) #data[,3:12]
957

```

```

958 # Estimacion de parámetros iniciales utilizando las distribuciones marginales
959 # SEVERIDAD
960 sd_alpha <- sqrt(diag(vcov(glm.Sev.RC)))
961 alpha_0 <- glm.Sev.RC.coefficients
962 delta_0 <- summary(glm.Sev.RC).dispersion
963 mu_0 <- exp(var_x%*%alpha_0)
964
965 # FRECUENCIA
966 beta_0 <- glm.Freq.RC.coefficients
967 sd_beta <- sqrt(diag(vcov(glm.Freq.RC)))
968 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
969
970 # REGRESION MEDIANTE COPULAS
971
972 # COPULA GUMBEL
973 family=5
974 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
975 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
  lambda_0)/(length(y)+1), family=family).par
976 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad-
  data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
977 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
978
979 # Estimación de las seudo-observaciones
980 u<-pgam(x,mu_0,delta_0)
981 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
982 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
983
984 # Creamos la función primera derivada de la Función Copula
985 Du_Frank <- function(u,v,theta){
986   u[u<=0]=0.0001
987   u[u>1]=1
988   v[v<=0]=0.0001
989   v[v>1]=1
990   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
991   #out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
992   #out <- (u^(-1)*exp((-log(u))^theta+(-log(v))^theta)^(1/theta))
993   out <- (exp(theta)*(exp(theta*v)-1))/(exp(theta*(u+1))+exp(theta*(v+1))-exp(theta)-exp(theta*(u+v)))
994   return(out)
995 }
996
997 # Función a maximizar
998 f_aux <- function(para){
999   D_u_0<-Du_Frank(u,v,para)-Du_Frank(u,vv,para)
1000   D_u_0[D_u_0<=0]=1
1001   D_u_0 <- log(D_u_0)
1002   out<-(-sum(D_u_0))
1003   return(out)
1004 }
1005
1006 # Proceso de maximización
1007 #para_ini <- theta2z(theta_ini,family = family)
1008 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
1009 #theta_0 <- z2theta(para_ifm,family = family)
1010 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
  para la regresión conjunta
1011 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
1012
1013 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
1014
1015 # Creamos nuestra función a maximizar
1016 #####f_aux_reg <- function(para){
1017 f_aux_reg <- function(para){
1018
1019   p <- ncol(var_x)
1020   q <- ncol(var_x)
1021   alpha <- para[1:p]
1022   beta <- para[(p+1):(p+q)]
1023   # theta <- z2theta(para[p+q+1],family)
1024   theta <- z2theta(para[p+q+1],family = family)
1025   delta <- para[p+q+2]
1026   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
1027   mu <- as.vector(exp(var_y%*%alpha))
1028   #dummy <- density_joint(x,y,mu,delta,theta,family=family,zt=FALSE)
1029   u2 <- pgam(data_model.Incurrido,mu,delta)
1030   v2<-ppois(y,lambda)

```

```

1031 vv2<-ppois(y=1,lambda)
1032 marginal.x <- dgam(data_model.Incurrido ,mu, delta)
1033 marginal.x[marginal.x>=1]=0
1034 marginal.x[marginal.x<=0]=0
1035 marginal.x[is.na(marginal.x)]=0
1036 marginal.x[marginal.x==Inf]=0
1037 marginal.x[marginal.x== -Inf]=0
1038 marginal.x[marginal.x==1Inf]=0
1039
1040 par_der <- Du_Frank(u2,v2,theta)
1041 par_der1 <- Du_Frank(u2, vv2, theta)
1042
1043 dummy<- par_der-par_der1
1044
1045 dummy[y==0]=par_der[y==0]
1046
1047 out<-marginal.x*dummy
1048
1049 out[out<=0]=1e-10
1050 ll <- -sum(log(out))
1051
1052 #if(negative==TRUE) ll <- (-ll)
1053 return(ll)
1054 }
1055
1056 ##########
1057 para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
1058 a<-now()
1059 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Estimación Maxima verosímil de los
1060 # parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
1061 b<-now()
1062 b-a #17,499 mins
1063 #load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
1064 #Parcial/JointModel_Gaus.RData")
1065 para_optim.value # Log_Likelihood
1066 vector <- para_optim.par
1067 p <- ncol(var_y)
1068 q <- ncol(var_y)
1069 Log_Likelihood<-para_optim.value
1070 # Parámetros optimizados
1071 alpha_frank<-vector[1:p]
1072 beta_frank<-vector[(p+1):(p+q)]
1073 theta_frank=z2theta(vector[p+q+1],family)
1074 delta_frank<-exp(vector[p+q+2]) #4.271691
1075 tau_frank<-BiCopPar2Tau(par=theta_frank,family=family)
1076 head(beta_0)
1077 head(beta_frank)
1078 head(alpha_0)
1079 head(alpha_frank)
1080
1081 # Determinamos los errores standar para el test de Wald
1082 hessian_frank <- para_optim.hessian
1083 Hinv <- ginv(hessian_frank)
1084 sd <- sqrt(diag(Hinv))
1085 sd.alpha_frank <-sd[1:p]
1086 sd.beta_frank <- sd[(p+1):(p+q)]
1087 sd.theta_frank <-sd[p+q+1]
1088
1089 # Ordenamos los coeficientes estimados en una tabla
1090 Reg_Cop_Frank <- data.frame(Estimate_Frec=round(beta_frank,4),Std.error.freq=round(sd.beta_frank,4),z_
1091 value.freq=round(beta_frank/sd.beta_frank,4),"Prob.freq"=round(2*(1-pnorm(abs(beta_frank/sd.beta_frank))))_
1092 ,4),Estimate_Sev=round(alpha_frank,4),Std.error.sev=round(sd.alpha_frank,4),z_value_sev=round(alpha_frank/
1093 sd.alpha_frank,4),"Prob.Sev"=round(2*(1-pnorm(abs(alpha_frank/sd.alpha_frank))),4))
1094 #Reg_Cop_Frank <- data.frame(Estimate_Frec=round(beta_frank,4),Std.error.freq=0,z_value.freq=0,"Prob.freq"
1095 "=0,Estimate.Sev=round(alpha_frank,4),Std.error.sev=0,z_value_sev=0,"Prob.Sev"=0)
1096 var_level<-row.names(Reg_Cop_Frank)
1097 Reg_Cop_Frank <- Reg_Cop_Frank %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
1098
1099 # Agregamos los interceptos
1100 temp <- tibble(
1101   var_level = c(
1102     "Año2022",
1103     "RC_Antiguedad.b. 2-9",
1104     "RC_Canal_VentaCorredores",
1105     "RC_Uso_VehParticular",
1106     "RC_Marca_VehL3",
1107     "RC_Marca_VehL4",
1108     "RC_Marca_VehL5",
1109     "RC_Marca_VehL6",
1110     "RC_Marca_VehL7",
1111     "RC_Marca_VehL8",
1112     "RC_Marca_VehL9",
1113     "RC_Marca_VehL10",
1114     "RC_Marca_VehL11",
1115     "RC_Marca_VehL12",
1116     "RC_Marca_VehL13",
1117     "RC_Marca_VehL14",
1118     "RC_Marca_VehL15",
1119     "RC_Marca_VehL16",
1120     "RC_Marca_VehL17",
1121     "RC_Marca_VehL18",
1122     "RC_Marca_VehL19",
1123     "RC_Marca_VehL20",
1124     "RC_Marca_VehL21",
1125     "RC_Marca_VehL22",
1126     "RC_Marca_VehL23",
1127     "RC_Marca_VehL24",
1128     "RC_Marca_VehL25",
1129     "RC_Marca_VehL26",
1130     "RC_Marca_VehL27",
1131     "RC_Marca_VehL28",
1132     "RC_Marca_VehL29",
1133     "RC_Marca_VehL30",
1134     "RC_Marca_VehL31",
1135     "RC_Marca_VehL32",
1136     "RC_Marca_VehL33",
1137     "RC_Marca_VehL34",
1138     "RC_Marca_VehL35",
1139     "RC_Marca_VehL36",
1140     "RC_Marca_VehL37",
1141     "RC_Marca_VehL38",
1142     "RC_Marca_VehL39",
1143     "RC_Marca_VehL40",
1144     "RC_Marca_VehL41",
1145     "RC_Marca_VehL42",
1146     "RC_Marca_VehL43",
1147     "RC_Marca_VehL44",
1148     "RC_Marca_VehL45",
1149     "RC_Marca_VehL46",
1150     "RC_Marca_VehL47",
1151     "RC_Marca_VehL48",
1152     "RC_Marca_VehL49",
1153     "RC_Marca_VehL50",
1154     "RC_Marca_VehL51",
1155     "RC_Marca_VehL52",
1156     "RC_Marca_VehL53",
1157     "RC_Marca_VehL54",
1158     "RC_Marca_VehL55",
1159     "RC_Marca_VehL56",
1160     "RC_Marca_VehL57",
1161     "RC_Marca_VehL58",
1162     "RC_Marca_VehL59",
1163     "RC_Marca_VehL60",
1164     "RC_Marca_VehL61",
1165     "RC_Marca_VehL62",
1166     "RC_Marca_VehL63",
1167     "RC_Marca_VehL64",
1168     "RC_Marca_VehL65",
1169     "RC_Marca_VehL66",
1170     "RC_Marca_VehL67",
1171     "RC_Marca_VehL68",
1172     "RC_Marca_VehL69",
1173     "RC_Marca_VehL70",
1174     "RC_Marca_VehL71",
1175     "RC_Marca_VehL72",
1176     "RC_Marca_VehL73",
1177     "RC_Marca_VehL74",
1178     "RC_Marca_VehL75",
1179     "RC_Marca_VehL76",
1180     "RC_Marca_VehL77",
1181     "RC_Marca_VehL78",
1182     "RC_Marca_VehL79",
1183     "RC_Marca_VehL80",
1184     "RC_Marca_VehL81",
1185     "RC_Marca_VehL82",
1186     "RC_Marca_VehL83",
1187     "RC_Marca_VehL84",
1188     "RC_Marca_VehL85",
1189     "RC_Marca_VehL86",
1190     "RC_Marca_VehL87",
1191     "RC_Marca_VehL88",
1192     "RC_Marca_VehL89",
1193     "RC_Marca_VehL90",
1194     "RC_Marca_VehL91",
1195     "RC_Marca_VehL92",
1196     "RC_Marca_VehL93",
1197     "RC_Marca_VehL94",
1198     "RC_Marca_VehL95",
1199     "RC_Marca_VehL96",
1200     "RC_Marca_VehL97",
1201     "RC_Marca_VehL98",
1202     "RC_Marca_VehL99",
1203     "RC_Marca_VehL100",
1204     "RC_Marca_VehL101",
1205     "RC_Marca_VehL102",
1206     "RC_Marca_VehL103",
1207     "RC_Marca_VehL104",
1208     "RC_Marca_VehL105",
1209     "RC_Marca_VehL106",
1210     "RC_Marca_VehL107",
1211     "RC_Marca_VehL108",
1212     "RC_Marca_VehL109",
1213     "RC_Marca_VehL110",
1214     "RC_Marca_VehL111",
1215     "RC_Marca_VehL112",
1216     "RC_Marca_VehL113",
1217     "RC_Marca_VehL114",
1218     "RC_Marca_VehL115",
1219     "RC_Marca_VehL116",
1220     "RC_Marca_VehL117",
1221     "RC_Marca_VehL118",
1222     "RC_Marca_VehL119",
1223     "RC_Marca_VehL120",
1224     "RC_Marca_VehL121",
1225     "RC_Marca_VehL122",
1226     "RC_Marca_VehL123",
1227     "RC_Marca_VehL124",
1228     "RC_Marca_VehL125",
1229     "RC_Marca_VehL126",
1230     "RC_Marca_VehL127",
1231     "RC_Marca_VehL128",
1232     "RC_Marca_VehL129",
1233     "RC_Marca_VehL130",
1234     "RC_Marca_VehL131",
1235     "RC_Marca_VehL132",
1236     "RC_Marca_VehL133",
1237     "RC_Marca_VehL134",
1238     "RC_Marca_VehL135",
1239     "RC_Marca_VehL136",
1240     "RC_Marca_VehL137",
1241     "RC_Marca_VehL138",
1242     "RC_Marca_VehL139",
1243     "RC_Marca_VehL140",
1244     "RC_Marca_VehL141",
1245     "RC_Marca_VehL142",
1246     "RC_Marca_VehL143",
1247     "RC_Marca_VehL144",
1248     "RC_Marca_VehL145",
1249     "RC_Marca_VehL146",
1250     "RC_Marca_VehL147",
1251     "RC_Marca_VehL148",
1252     "RC_Marca_VehL149",
1253     "RC_Marca_VehL150",
1254     "RC_Marca_VehL151",
1255     "RC_Marca_VehL152",
1256     "RC_Marca_VehL153",
1257     "RC_Marca_VehL154",
1258     "RC_Marca_VehL155",
1259     "RC_Marca_VehL156",
1260     "RC_Marca_VehL157",
1261     "RC_Marca_VehL158",
1262     "RC_Marca_VehL159",
1263     "RC_Marca_VehL160",
1264     "RC_Marca_VehL161",
1265     "RC_Marca_VehL162",
1266     "RC_Marca_VehL163",
1267     "RC_Marca_VehL164",
1268     "RC_Marca_VehL165",
1269     "RC_Marca_VehL166",
1270     "RC_Marca_VehL167",
1271     "RC_Marca_VehL168",
1272     "RC_Marca_VehL169",
1273     "RC_Marca_VehL170",
1274     "RC_Marca_VehL171",
1275     "RC_Marca_VehL172",
1276     "RC_Marca_VehL173",
1277     "RC_Marca_VehL174",
1278     "RC_Marca_VehL175",
1279     "RC_Marca_VehL176",
1280     "RC_Marca_VehL177",
1281     "RC_Marca_VehL178",
1282     "RC_Marca_VehL179",
1283     "RC_Marca_VehL180",
1284     "RC_Marca_VehL181",
1285     "RC_Marca_VehL182",
1286     "RC_Marca_VehL183",
1287     "RC_Marca_VehL184",
1288     "RC_Marca_VehL185",
1289     "RC_Marca_VehL186",
1290     "RC_Marca_VehL187",
1291     "RC_Marca_VehL188",
1292     "RC_Marca_VehL189",
1293     "RC_Marca_VehL190",
1294     "RC_Marca_VehL191",
1295     "RC_Marca_VehL192",
1296     "RC_Marca_VehL193",
1297     "RC_Marca_VehL194",
1298     "RC_Marca_VehL195",
1299     "RC_Marca_VehL196",
1300     "RC_Marca_VehL197",
1301     "RC_Marca_VehL198",
1302     "RC_Marca_VehL199",
1303     "RC_Marca_VehL200",
1304     "RC_Marca_VehL201",
1305     "RC_Marca_VehL202",
1306     "RC_Marca_VehL203",
1307     "RC_Marca_VehL204",
1308     "RC_Marca_VehL205",
1309     "RC_Marca_VehL206",
1310     "RC_Marca_VehL207",
1311     "RC_Marca_VehL208",
1312     "RC_Marca_VehL209",
1313     "RC_Marca_VehL210",
1314     "RC_Marca_VehL211",
1315     "RC_Marca_VehL212",
1316     "RC_Marca_VehL213",
1317     "RC_Marca_VehL214",
1318     "RC_Marca_VehL215",
1319     "RC_Marca_VehL216",
1320     "RC_Marca_VehL217",
1321     "RC_Marca_VehL218",
1322     "RC_Marca_VehL219",
1323     "RC_Marca_VehL220",
1324     "RC_Marca_VehL221",
1325     "RC_Marca_VehL222",
1326     "RC_Marca_VehL223",
1327     "RC_Marca_VehL224",
1328     "RC_Marca_VehL225",
1329     "RC_Marca_VehL226",
1330     "RC_Marca_VehL227",
1331     "RC_Marca_VehL228",
1332     "RC_Marca_VehL229",
1333     "RC_Marca_VehL230",
1334     "RC_Marca_VehL231",
1335     "RC_Marca_VehL232",
1336     "RC_Marca_VehL233",
1337     "RC_Marca_VehL234",
1338     "RC_Marca_VehL235",
1339     "RC_Marca_VehL236",
1340     "RC_Marca_VehL237",
1341     "RC_Marca_VehL238",
1342     "RC_Marca_VehL239",
1343     "RC_Marca_VehL240",
1344     "RC_Marca_VehL241",
1345     "RC_Marca_VehL242",
1346     "RC_Marca_VehL243",
1347     "RC_Marca_VehL244",
1348     "RC_Marca_VehL245",
1349     "RC_Marca_VehL246",
1350     "RC_Marca_VehL247",
1351     "RC_Marca_VehL248",
1352     "RC_Marca_VehL249",
1353     "RC_Marca_VehL250",
1354     "RC_Marca_VehL251",
1355     "RC_Marca_VehL252",
1356     "RC_Marca_VehL253",
1357     "RC_Marca_VehL254",
1358     "RC_Marca_VehL255",
1359     "RC_Marca_VehL256",
1360     "RC_Marca_VehL257",
1361     "RC_Marca_VehL258",
1362     "RC_Marca_VehL259",
1363     "RC_Marca_VehL260",
1364     "RC_Marca_VehL261",
1365     "RC_Marca_VehL262",
1366     "RC_Marca_VehL263",
1367     "RC_Marca_VehL264",
1368     "RC_Marca_VehL265",
1369     "RC_Marca_VehL266",
1370     "RC_Marca_VehL267",
1371     "RC_Marca_VehL268",
1372     "RC_Marca_VehL269",
1373     "RC_Marca_VehL270",
1374     "RC_Marca_VehL271",
1375     "RC_Marca_VehL272",
1376     "RC_Marca_VehL273",
1377     "RC_Marca_VehL274",
1378     "RC_Marca_VehL275",
1379     "RC_Marca_VehL276",
1380     "RC_Marca_VehL277",
1381     "RC_Marca_VehL278",
1382     "RC_Marca_VehL279",
1383     "RC_Marca_VehL280",
1384     "RC_Marca_VehL281",
1385     "RC_Marca_VehL282",
1386     "RC_Marca_VehL283",
1387     "RC_Marca_VehL284",
1388     "RC_Marca_VehL285",
1389     "RC_Marca_VehL286",
1390     "RC_Marca_VehL287",
1391     "RC_Marca_VehL288",
1392     "RC_Marca_VehL289",
1393     "RC_Marca_VehL290",
1394     "RC_Marca_VehL291",
1395     "RC_Marca_VehL292",
1396     "RC_Marca_VehL293",
1397     "RC_Marca_VehL294",
1398     "RC_Marca_VehL295",
1399     "RC_Marca_VehL296",
1400     "RC_Marca_VehL297",
1401     "RC_Marca_VehL298",
1402     "RC_Marca_VehL299",
1403     "RC_Marca_VehL300",
1404     "RC_Marca_VehL301",
1405     "RC_Marca_VehL302",
1406     "RC_Marca_VehL303",
1407     "RC_Marca_VehL304",
1408     "RC_Marca_VehL305",
1409     "RC_Marca_VehL306",
1410     "RC_Marca_VehL307",
1411     "RC_Marca_VehL308",
1412     "RC_Marca_VehL309",
1413     "RC_Marca_VehL310",
1414     "RC_Marca_VehL311",
1415     "RC_Marca_VehL312",
1416     "RC_Marca_VehL313",
1417     "RC_Marca_VehL314",
1418     "RC_Marca_VehL315",
1419     "RC_Marca_VehL316",
1420     "RC_Marca_VehL317",
1421     "RC_Marca_VehL318",
1422     "RC_Marca_VehL319",
1423     "RC_Marca_VehL320",
1424     "RC_Marca_VehL321",
1425     "RC_Marca_VehL322",
1426     "RC_Marca_VehL323",
1427     "RC_Marca_VehL324",
1428     "RC_Marca_VehL325",
1429     "RC_Marca_VehL326",
1430     "RC_Marca_VehL327",
1431     "RC_Marca_VehL328",
1432     "RC_Marca_VehL329",
1433     "RC_Marca_VehL330",
1434     "RC_Marca_VehL331",
1435     "RC_Marca_VehL332",
1436     "RC_Marca_VehL333",
1437     "RC_Marca_VehL334",
1438     "RC_Marca_VehL335",
1439     "RC_Marca_VehL336",
1440     "RC_Marca_VehL337",
1441     "RC_Marca_VehL338",
1442     "RC_Marca_VehL339",
1443     "RC_Marca_VehL340",
1444     "RC_Marca_VehL341",
1445     "RC_Marca_VehL342",
1446     "RC_Marca_VehL343",
1447     "RC_Marca_VehL344",
1448     "RC_Marca_VehL345",
1449     "RC_Marca_VehL346",
1450     "RC_Marca_VehL347",
1451     "RC_Marca_VehL348",
1452     "RC_Marca_VehL349",
1453     "RC_Marca_VehL350",
1454     "RC_Marca_VehL351",
1455     "RC_Marca_VehL352",
1456     "RC_Marca_VehL353",
1457     "RC_Marca_VehL354",
1458     "RC_Marca_VehL355",
1459     "RC_Marca_VehL356",
1460     "RC_Marca_VehL357",
1461     "RC_Marca_VehL358",
1462     "RC_Marca_VehL359",
1463     "RC_Marca_VehL360",
1464     "RC_Marca_VehL361",
1465     "RC_Marca_VehL362",
1466     "RC_Marca_VehL363",
1467     "RC_Marca_VehL364",
1468     "RC_Marca_VehL365",
1469     "RC_Marca_VehL366",
1470     "RC_Marca_VehL367",
1471     "RC_Marca_VehL368",
1472     "RC_Marca_VehL369",
1473     "RC_Marca_VehL370",
1474     "RC_Marca_VehL371",
1475     "RC_Marca_VehL372",
1476     "RC_Marca_VehL373",
1477     "RC_Marca_VehL374",
1478     "RC_Marca_VehL375",
1479     "RC_Marca_VehL376",
1480     "RC_Marca_VehL377",
1481     "RC_Marca_VehL378",
1482     "RC_Marca_VehL379",
1483     "RC_Marca_VehL380",
1484     "RC_Marca_VehL381",
1485     "RC_Marca_VehL382",
1486     "RC_Marca_VehL383",
1487     "RC_Marca_VehL384",
1488     "RC_Marca_VehL385",
1489     "RC_Marca_VehL386",
1490     "RC_Marca_VehL387",
1491     "RC_Marca_VehL388",
1492     "RC_Marca_VehL389",
1493     "RC_Marca_VehL390",
1494     "RC_Marca_VehL391",
1495     "RC_Marca_VehL392",
1496     "RC_Marca_VehL393",
1497     "RC_Marca_VehL394",
1498     "RC_Marca_VehL395",
1499     "RC_Marca_VehL396",
1500     "RC_Marca_VehL397",
1501     "RC_Marca_VehL398",
1502     "RC_Marca_VehL399",
1503     "RC_Marca_VehL400",
1504     "RC_Marca_VehL401",
1505     "RC_Marca_VehL402",
1506     "RC_Marca_VehL403",
1507     "RC_Marca_VehL404",
1508     "RC_Marca_VehL405",
1509     "RC_Marca_VehL406",
1510     "RC_Marca_VehL407",
1511     "RC_Marca_VehL408",
1512     "RC_Marca_VehL409",
1513     "RC_Marca_VehL410",
1514     "RC_Marca_VehL411",
1515     "RC_Marca_VehL412",
1516     "RC_Marca_VehL413",
1517     "RC_Marca_VehL414",
1518     "RC_Marca_VehL415",
1519     "RC_Marca_VehL416",
1520     "RC_Marca_VehL417",
1521     "RC_Marca_VehL418",
1522     "RC_Marca_VehL419",
1523     "RC_Marca_VehL420",
1524     "RC_Marca_VehL421",
1525     "RC_Marca_VehL422",
1526     "RC_Marca_VehL423",
1527     "RC_Marca_VehL424",
1528     "RC_Marca_VehL425",
1529     "RC_Marca_VehL426",
1530     "RC_Marca_VehL427",
1531     "RC_Marca_VehL428",
1532     "RC_Marca_VehL429",
1533     "RC_Marca_VehL430",
1534     "RC_Marca_VehL431",
1535     "RC_Marca_VehL432",
1536     "RC_Marca_VehL433",
1537     "RC_Marca_VehL434",
1538     "RC_Marca_VehL435",
1539     "RC_Marca_VehL436",
1540     "RC_Marca_VehL437",
1541     "RC_Marca_VehL438",
1542     "RC_Marca_VehL439",
1543     "RC_Marca_VehL440",
1544     "RC_Marca_VehL441",
1545     "RC_Marca_VehL442",
1546     "RC_Marca_VehL443",
1547     "RC_Marca_VehL444",
1548     "RC_Marca_VehL445",
1549     "RC_Marca_VehL446",
1550     "RC_Marca_VehL447",
1551     "RC_Marca_VehL448",
1552     "RC_Marca_VehL449",
1553     "RC_Marca_VehL450",
1554     "RC_Marca_VehL451",
1555     "RC_Marca_VehL452",
1556     "RC_Marca_VehL453",
1557     "RC_Marca_VehL454",
1558     "RC_Marca_VehL455",
1559     "RC_Marca_VehL456",
1560     "RC_Marca_VehL457",
1561     "RC_Marca_VehL458",
1562     "RC_Marca_VehL459",
1563     "RC_Marca_VehL460",
1564     "RC_Marca_VehL461",
1565     "RC_Marca_VehL462",
1566     "RC_Marca_VehL463",
1567     "RC_Marca_VehL464",
1568     "RC_Marca_VehL465",
1569     "RC_Marca_VehL466",
1570     "RC_Marca_VehL467",
1571     "RC_Marca_VehL468",
1572     "RC_Marca_VehL469",
1573     "RC_Marca_VehL470",
1574     "RC_Marca_VehL471",
1575     "RC_Marca_VehL472",
1576     "RC_Marca_VehL473",
1577     "RC_Marca_VehL474",
1578     "RC_Marca_VehL475",
1579     "RC_Marca_VehL476",
1580     "RC_Marca_VehL477",
1581     "RC_Marca_VehL478",
1582     "RC_Marca_VehL479",
1583     "RC_Marca_VehL480",
1584     "RC_Marca_VehL481",
1585     "RC_Marca_VehL482",
1586     "RC_Marca_VehL483",
1587     "RC_Marca_VehL484",
1588     "RC_Marca_VehL485",
1589     "RC_Marca_VehL486",
1590     "RC_Marca_VehL487",
1591     "RC_Marca_VehL488",
1592     "RC_Marca_VehL489",
1593     "RC_Marca_VehL490",
1594     "RC_Marca_VehL491",
1595     "RC_Marca_VehL492",
1596     "RC_Marca_VehL493",
1597     "RC_Marca_VehL494",
1598     "RC_Marca_VehL495",
1599     "RC_Marca_VehL496",
1600     "RC_Marca_VehL497",
1601     "RC_Marca_VehL498",
1602     "RC_Marca_VehL499",
1603     "RC_Marca_VehL500",
1604     "RC_Marca_VehL501",
1605     "RC_Marca_VehL502",
1606     "RC_Marca_VehL503",
1607     "RC_Marca_VehL504",
1608     "RC_Marca_VehL505",
1609     "RC_Marca_VehL506",
1610     "RC_Marca_VehL507",
1611     "RC_Marca_VehL508",
1612     "RC_Marca_VehL509",
1613     "RC_Marca_VehL510",
1614     "RC_Marca_VehL511",
1615     "RC_Marca_VehL512",
1616     "RC_Marca_VehL513",
1617     "RC_Marca_VehL514",
1618     "RC_Marca_VehL515",
1619     "RC_Marca_VehL516",
1620     "RC_Marca_VehL517",
1621     "RC_Marca_VehL518",
1622     "RC_Marca_VehL519",
1623     "RC_Marca_VehL520",
1624     "RC_Marca_VehL521",
1625     "RC_Marca_VehL522",
1626     "RC_Marca_VehL523",
1627     "RC_Marca_VehL524",
1628     "RC_Marca_VehL525",
1629     "RC_Marca_VehL526",
1630     "RC_Marca_VehL527",
1631     "RC_Marca_VehL528",
1632     "RC_Marca_VehL529",
1633     "RC_Marca_VehL530",
1634     "RC_Marca_VehL531",
1635     "RC_Marca_VehL532",
1636     "RC_Marca_VehL533",
1637     "RC_Marca_VehL534",
1638     "RC_Marca_VehL535",
1639     "RC_Marca_VehL536",
1640     "RC_Marca_VehL537",
1641     "RC_Marca_VehL538",
1642     "RC_Marca_VehL539",
1643     "RC_Marca_VehL540",
1644     "RC_Marca_VehL541",
1645     "RC_Marca_VehL542",
1646     "RC_Marca_VehL543",
1647     "RC_Marca_VehL544",
1648     "RC_Marca_VehL545",
1649     "RC_Marca_VehL546",
1650     "RC_Marca_VehL547",
1651     "RC_Marca_VehL548",
1652     "RC_Marca_VehL549",
1653     "RC_Marca_VehL550",
1654     "RC_Marca_VehL551",
1655     "RC_Marca_VehL552",
1656     "RC_Marca_VehL553",
1657     "RC_Marca_VehL554",
1658     "RC_Marca_VehL555",
1659     "RC_Marca_VehL556",
1660     "RC_Marca_VehL557",
1661     "RC_Marca_VehL558",
1662     "RC_Marca_VehL559",
1663     "RC_Marca_VehL560",
1664     "RC_Marca_VehL561",
1665     "RC_Marca_VehL562",
1666     "RC_Marca_VehL563",
1667     "RC_Marca_VehL564",
1668     "RC_Marca_VehL565",
1669     "RC_Marca_VehL566",
1670     "RC_Marca_VehL567",
1671     "RC_Marca_VehL568",
1672     "RC_Marca_VehL569",
1673     "RC_Marca_VehL570",
1674     "RC_Marca_VehL571",
1675     "RC_Marca_VehL572",
1676     "RC_Marca_VehL573",
1677     "RC_Marca_VehL574",
1678     "RC_Marca_VehL575",
1679     "RC_Marca_VehL576",
1680     "RC_Marca_VehL577",
1681     "RC_Marca_VehL578",
1682     "RC_Marca_VehL579",
1683     "RC_Marca_VehL580",
1684     "RC_Marca_VehL581",
1685     "RC_Marca_VehL582",
1686     "RC_Marca_VehL583",
1687     "RC_Marca_VehL584",
1688     "RC_Marca_VehL585",
1689     "RC_Marca_VehL586",
169
```

```

1101      "RC_ZonaA2",
1102      "RC_Edad43-52"
1103    ),
1104    Estimate_Frec = 0
1105  ,
1106  Estimate_Sev = 0
1107 )
1108
1109 # coeficientes del modelo de regresion con dependencia
1110 Reg_Cop_Frank <- Reg_Cop_Frank %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)`=
1111   round(exp(Estimate_Frec),4), `exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
1112 Reg_Cop_Frank <-Reg_Cop_Frank %>% relocate(`exp(Estimate_Frec)`, .after=Prob_frec)
1113
1114 # Guardamos el modelo conjunto y los coeficientes
1115 write.csv(Reg_Cop_Frank, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
1116 GLM_Conjuntos/Responsabilidad_Civil/Coef_Frank.csv", fileEncoding = "Latin1")
1117 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
1118 Conjuntos/Responsabilidad_Civil/JointModel_Frank.RData")
1119
1120 # Simulacion de la Pérdida Total
1121 # =====
1122 # CON DEPENDENCIA
1123 # -----
1124 lambda_loss_dep <- exp(var_x%*%beta_frank)*data_CM.Expuesto
1125 mu_loss_dep <- exp(var_x%*%alpha_frank)
1126 delta_loss_dep <- delta_frank
1127
1128
1129 # Función de densidad fx
1130
1131 f_loss <-
1132   function(loss ,mu,d, lam ,theta){
1133     n<-length(loss)
1134     if (length(lam)==1) lam <- rep(lam,n)
1135     if (length(mu)==1) mu <- rep(mu,n)
1136     out <- vector(length=n)
1137
1138     for(i in 1:n){
1139       N <-1:20
1140       v <-ppois(N, lam[i])
1141       vv <-ppois(N-1, lam[i])
1142       u <-pgam(loss[i]/N,mu[i],d)
1143
1144       Der_cop <-Du_Frank(u,v,theta)-Du_Frank(u,vv,theta) # Primera derivada de la función copula
1145       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N # Teorela del Total Loss
1146       out[i] <-sum(dummy)
1147     }
1148
1149     out[loss <=0]=0
1150     return(out)
1151   }
1152
1153 # Función de densidad acumulada Fx
1154 F_loss <-
1155   function(loss ,mu,d, lam ,theta){
1156     out<-vector(length=length(loss))
1157
1158     for(i in 1:length(loss)){
1159       floss <- function(s){
1160         f_loss(s,mu[i],d, lam[i],theta)
1161       }
1162       out[i] <- integrate(floss ,0 ,loss[i]).value # Integral para el cálculo del a F(x) a partir de f(x)
1163     }
1164     return(out)
1165   }
1166
1167 # Estimación del Total Loss (fuerte dedicacion computacional)
1168
1169 #k <-length(lambda_loss_dep)
1170 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
1171 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
1172 L <- vector(length=k)
1173 S <- vector(length=m)

```

```

1174
1175   for(j in 1:m){
1176     r_uni<-runif(k)
1177     for(i in 1:k){
1178       f_root <function(s){
1179         F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_frank)-r_uni[i]
1180       }
1181     }
1182     tryCatch(
1183       error = function(cond) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
1184       loss<-uniroot(f_root,lower = 0,upper = 500000).root
1185     )
1186     #print(loss.root)
1187     L[i]<-loss
1188     perc<-paste(i/k*100,"%")
1189     print(perc)
1190   }
1191   S[j]<-sum(L)
1192   perc<-paste(j/m*100,"%")
1193   print(perc)
1194   #print(S[j])
1195 }
1196 }
1197 S
1198
1199 # CON INDEPENDENCIA
1200 #
1201 lambda_loss <-predict(glm.Frec.RC, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO
1202 mu_loss      <-predict(glm.Sev.RC, newdata=data_CM, type='response')
1203 delta_loss   <-delta_0
1204
1205 M<-seq(1:10000)
1206 for(i in 1:10000){
1207   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
1208   S<-vector(length=length(lambda_loss))
1209   for(j in 1:length(muestra_cant)){
1210     if(muestra_cant[j]==0){
1211       S[j]=0
1212     } else{
1213       S[j]=sum(rgam(n=muestra_cant[j],mu_loss[j],delta_loss))
1214     }
1215   }
1216
1217   M[i] <- sum(S)
1218   porc<-paste((i/10000)*100,"%")
1219   print(porc)
1220   #print(M[i])
1221 }
1222
1223 hist(M,breaks = 50)
1224 mean(M)
1225 sum(data_model.Incurrido)
1226
1227 # Simulacion de la Frecuencia con Independencia
1228 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
1229 # Frecuencia
1230 N<-seq(1:10000)
1231 for(i in 1:10000){
1232   N[i] <- sum(simulate(glm.Frec.RC,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
1233   porc<-paste((i/10000)*100,"%")
1234   print(porc)
1235 }
1236
1237 hist(N,breaks = 50)
1238 mean(N)
1239 sum(data_model.Cantidad)/sum(data_model.Expuesto)

```

C.6. Modelo Conjunto - Asistencias

```

1 #####
2 #####
3 #####
4 ##### REGRESION BASADO EN COPULAS #####
5 ##### ASISTENCIA #####

```

```

6 ##### #####
7 ##### #####
8 ##### #####
9 ##### #####
10 # Cargamos las librerías necesarias #####
11
12 #library(copula)
13 library(RODBC)
14 library(tidyverse)
15 library(MASS)
16 #library(GJRM)
17 #library(devtools)
18 library(CopulaRegression)
19 library(VineCopula)
20 library(optimx)
21 #install_url('http://cran.r-project.org/src/contrib/Archive/CopulaRegression/CopulaRegression_0.1-5.tar.gz')
22
23 ##### #####
24 # REGRESION GAUSSIANA #####
25 ##### #####
26
27 # Set Directory
28 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia")
29
30 # Carga los modelos GLM MARGINALES calculados en otros querys
31 load("glm.Frec.AS.RData")
32 load("glm.Sev.AS.RData")
33 load("Data_AS.RData")
34 load("Data_AS_model.RData")
35 load("Data_AS_model_CM.RData")
36
37 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
38 # SEVERIDAD
39 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
40 var_x <- model.matrix(glm.Sev.AS) #data_CM[,3:12]
41 # FRECUENCIA
42 y <- data_model.Cantidad
43 Expuestos <- data_model.Expuesto
44 var_y <- model.matrix(glm.Frec.AS) #data[,3:12]
45
46 # Estimacion de parámetros iniciales utilizando las distribuciones marginales
47 # SEVERIDAD
48 sd_alpha <- sqrt(diag(vcov(glm.Sev.AS)))
49 alpha_0 <- glm.Sev.AS.coefficients
50 delta_0 <- summary(glm.Sev.AS).dispersion
51 mu_0 <- exp(var_x%*%alpha_0)
52
53 # FRECUENCIA
54 beta_0 <- glm.Frec.AS.coefficients
55 sd_beta <- sqrt(diag(vcov(glm.Frec.AS)))
56 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
57
58 # REGRESION MEDIANTE COPULAS
59
60 # COPULA GAUSSIANA
61 family=1
62 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
63 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-lambda_0)/(length(y)+1), family=family).par
64 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
65 tau_ini <- BiCopar2Tau(par=theta_ini, family = family)
66
67 # Estimación de las seudo-observaciones
68 u<-pgam(x,mu_0,delta_0)
69 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
70 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
71
72 # Creamos la función primera derivada de la Función Copula
73 Du_Gaus <- function(u,v,theta){
74   u[u<=0]=0.001
75   u[u>=1]=0.999
76   v[v<=0]=0.001
77   v[v>=1]=0.999
78   out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
79   return(out)

```

```

80      }
81
82  # Función a maximizar
83  f_aux <- function(para){
84    Du_0<-Du_Gaus(u,v,para)-Du_Gaus(u,vv,para)
85    Du_0[Du_0<=0]=1
86    Du_0 <- log(Du_0)
87    out<-(-sum(Du_0))
88    return(out)
89  }
90
91  # Proceso de maximización
92  #para_ini <- theta2z(theta_ini,family = family)
93  para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
94  #theta_0 <- z2theta(para_ifm,family = family)
95  theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
96  # para la regresión conjunta
97  tau <- BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
98
99  # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
100
101 # Creamos nuestra función a maximizar
102 #####f_aux_reg <- function(para){
103
104   p      <- ncol(var_x)
105   q      <- ncol(var_x)
106   alpha <- para[1:p]
107   beta  <- para[(p+1):(p+q)]
108   # theta <- z2theta(para[p+q+1],family)
109   theta <- para[p+q+1]
110   delta <- para[p+q+2]
111   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
112   mu    <- as.vector(exp(var_y%*%alpha))
113   #dummy <- density_joint(x,y,mu,delta,lambda,theta,family=family,zt=FALSE)
114   u2 <- pgam(data_model.Incurrido,mu,delta)
115   v2<-ppois(y,lambda)
116   vv2<-ppois(y-1,lambda)
117
118   marginal.x <- dgam(data_model.Incurrido,mu,delta)
119   marginal.x[marginal.x>=1]=0
120   marginal.x[marginal.x<=0]=0
121   marginal.x[is.na(marginal.x)]=0
122   marginal.x[marginal.x==Inf]=0
123   marginal.x[marginal.x==Inf]=0
124
125   par_der <- Du_Gaus(u2,v2,theta)
126   par_der1 <- Du_Gaus(u2,vv2,theta)
127
128   dummy<- par_der-par_der1
129
130   dummy[y==0]=par_der[y==0]
131
132   out<-marginal.x*dummy
133
134   out[out<=0]=1e-10
135   ll     <- -sum(log(out))
136
137   #if(negative==TRUE) ll <- (-ll)
138   return(ll)
139 }
140
141 #####
142
143 para_0<-c(alpha_0,beta_0,theta_0,delta_0) # Parámetros iniciales
144 a<-now()
145 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Demora 6min
146 b<-now()
147 b-a
148 #load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_Parcial/JointModel_Gaus.RData")
149 para_optim.value # Log_Likelihood
150 vector <- para_optim.par
151 p       <- ncol(var_y)
152 q       <- ncol(var_y)
153

```

```

154 # Parámetros optimizados
155 alpha_gau<-vector[1:p]
156 beta_gau<-vector[(p+1):(p+q)]
157 theta_gau=z2theta(vector[p+q+1],family)
158 delta_gau<-vector[p+q+2]
159 tau_gau<-BiCopPar2Tau(par=theta_gau,family=family)
160 head(beta_0)
161 head(beta_gau)
162 head(alpha_0)
163 head(alpha_gau)
164 # Determinamos los errores estandar para el test de Wald
165 hessian_gau <- para_optim.hessian
166 Hinv <- ginv(hessian_gau)
167 sd <- sqrt(diag(Hinv))
168 sd.alpha_gau <-sd[1:p]
169 sd.beta_gau <- sd[(p+1):(p+q)]
170 sd.theta_gau <-sd[p+q+1]
171
172 # Ordenamos los coeficientes estimados en una tabla
173 Reg_Cop_Gausiana <- data.frame(Estimate_Frec=round(beta_gau,4),Std.error.freq=round(sd.beta_gau,4),z_value_frec=round(beta_gau/sd.beta_gau,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_gau/sd.beta_gau))),4),
Estimate_Sev=round(alpha_gau,4),Std.error.sev=round(sd.alpha_gau,4),z_value_sev=round(alpha_gau/sd.alpha_gau,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_gau/sd.alpha_gau))),4))
#Reg_Cop_Gausiana <- data.frame(Estimate_Frec=round(beta_gau,4),Std.error.freq=0,z_value_frec=0,"Prob_frec"=0,Estimate_Sev=round(alpha_gau,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
174 var_level<-row.names(Reg_Cop_Gausiana)
175 Reg_Cop_Gausiana <- Reg_Cop_Gausiana %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
176
177 # Agregamos los interceptos
178 temp <- tibble(
179   var_level = c(
180     "Año2022",
181     "AS_Antiguedad_b_1-2",
182     "AS_Canal_VentaCorredores",
183     "AS_Uso_Vehicular",
184     "AS_Marca_VehL2",
185     "AS_ZonaA3",
186     "AS_SAa.[1000-10000]>"
187   ),
188   Estimate_Frec = 0
189   ,
190   Estimate_Sev = 0
191 )
192
193
194 # coeficientes del modelo de regresion con dependencia
195 Reg_Cop_Gausiana <- Reg_Cop_Gausiana %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)`=round(exp(Estimate_Frec),4), `exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
196
197 Reg_Cop_Gausiana <-Reg_Cop_Gausiana %>% relocate(`exp(Estimate_Frec)`, .after=Prob_frec)
198
199 # Guardamos el modelo conjunto y los coeficientes
200 write.csv(Reg_Cop_Gausiana, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Asistencia/Coeff_Gaus.csv", fileEncoding = "Latin1")
201 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Asistencia/JointModel_Gaus.RData")
202
203 # Simulacion de la Pérdida Total
204 # =====
205
206 # CON DEPENDENCIA
207 # -----
208 lambda_loss_dep <- exp(var_x%*%beta_gau)*data_CM.Expuesto
209 mu_loss_dep <- exp(var_x%*%alpha_gau)
210 delta_loss_dep <- delta_gau
211
212
213 # Función de densidad fx
214
215 f_loss <-
216   function(loss, mu, d, lam, theta){
217     n<-length(loss)
218     if (length(lam)==1) lam <- rep(lam,n)
219     if (length(mu)==1) mu <- rep(mu,n)
220     out <- vector(length=n)
221

```

```

222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
for(i in 1:n){
  N <-1:20
  v <-ppois(N, lam[i])
  vv <-ppois(N-1, lam[i])
  u <-pgam(loss[i]/N, mu[i], d)

  Der_cop <-Du_Gaus(u, v, theta)-Du_Gaus(u, vv, theta)
  dummy <-Der_cop*dgam(loss[i]/N, mu[i], d)/N
  out[i] <-sum(dummy)
}

out[loss<=0]=0
return(out)
}

# Función de densidad acumulada Fx
F_loss <-
function(loss, mu, d, lam, theta){
  out<-vector(length=length(loss))

  for(i in 1:length(loss)){
    floss <- function(s){
      f_loss(s, mu[i], d, lam[i], theta)
    }
    out[i] <- integrate(floss, 0, loss[i]).value
  }
  return(out)
}

# Estimación del Total Loss (fuerte dedicacion computacional)
#k <-length(lambda_loss_dep)
k <- length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
m <- 100 # Cantidad de muestras aleatorias (Deben ser 10 mil)
L <- vector(length=k)
S <- vector(length=m)

for(j in 1:m){
  r_uni<-runif(k)
  for(i in 1:k){
    f_root <-function(s){
      F_loss(s, mu_loss_dep[i], delta_loss_dep, lambda_loss_dep[i], theta_gau)-r_uni[i]
    }

    tryCatch(
      error = function(cond) mu_loss_dep[i]*lambda_loss_dep[i],
      loss<-uniroot(f_root, lower = 0, upper = 500000)
    )

    #print(loss.root)
    L[i]<-loss.root
    perc<-paste(i/k*100, "%")
    print(perc)
  }

  S[j]<-sum(L)
  perc<-paste(j/m*100, "%")
  print(perc)
  #print(S[j])
}

head(S)

# CON INDEPENDENCIA
# -----
lambda_loss <-predict(glm.Frec.AS, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN
MODELO DE RIESGO COLECTIVO
mu_loss <-predict(glm.Sev.AS, newdata=data_CM, type='response')
delta_loss <-delta_0

M<-seq(1:10000)
for(i in 1:10000){
  muestra_cant<-rpois(length(lambda_loss), lambda_loss)
  S<-vector(length=length(lambda_loss))
  for(j in 1:length(muestra_cant)){
    if(muestra_cant[j]==0){
      S[j]=0
    } else {

```

```

297           S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
298       }
299   }
300
301   M[i] <- sum(S)
302   porc<-paste((i/10000)*100,"%")
303   print(porc)
304   #print(M[i])
305   }
306
307   hist(M,breaks = 50)
308   mean(M)
309   sum(data.Incurrido)
310
311
312 # Simulacion de la Frecuencia con Independencia
313 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
314 # Frecuencia
315 N<-seq(1:10000)
316 for(i in 1:10000){
317   N[i] <- sum(simulate(glm.Frec.AS,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
318   porc<-paste((i/10000)*100,"%")
319   print(porc)
320   }
321   hist(N,breaks = 50)
322   mean(N)
323   sum(data.Frecuencia)/sum(data.Expuesto)
324
325 ##########
326 # Limpiamos el ambiente de variables
327 rm(list = ls())
328 gc()
329
330 #####
331 # REGRESION CLAYTON
332 #####
333
334 # Set Directory
335 setwd("C:/Users/josephgarcial/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia")
336
337 # Carga los modelos GLM MARGINALES calculados en otros querys
338 load("glm.Frec.AS.RData")
339 load("glm.Sev.AS.RData")
340 load("Data_AS.RData")
341 load("Data_AS_model.RData")
342 load("Data_AS_model_CM.RData")
343
344 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
345 # SEVERIDAD
346 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
347 var_x <- model.matrix(glm.Sev.AS) #data_CM[,3:12]
348 # FRECUENCIA
349 y <- data_model.Cantidad
350 Expuestos <- data_model.Expuesto
351 var_y <- model.matrix(glm.Frec.AS) #data[,3:12]
352
353 # Estimacion de parametros iniciales utilizando las distribuciones marginales
354 # SEVERIDAD
355 sd_alpha <- sqrt(diag(vcov(glm.Sev.AS)))
356 alpha_0 <- glm.Sev.AS.coefficients
357 delta_0 <- summary(glm.Sev.AS).dispersion
358 mu_0 <- exp(var_x%*%alpha_0)
359
360 # FRECUENCIA
361 beta_0 <- glm.Frec.AS.coefficients
362 sd_beta <- sqrt(diag(vcov(glm.Frec.AS)))
363 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
364
365 # REGRESION MEDIANTE COPULAS
366
367 # COPULA CLAYTON
368 family=3
369 # Estimacion inicial del parametro theta para estimarlo mediante el algoritmo de MPLE
370 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
lambda_0)/(length(y)+1), family=family).par

```

```

371 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*
372 data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
373 tau_ini <- BiCopPar2Tau(par=theta_ini ,family = family)
374
375 # Estimación de las seudo-observaciones
376 u<-pgam(x,mu_0,delta_0)
377 v <- ppois(data_CM.Cantidad ,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
378 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
379
380 # Creamos la función primera derivada de la Función Copula
381 Du_Clay <- function(u,v,theta){
382   u[u<=0]=0.0001
383   u[u>1]=1
384   v[v<=0]=0.0001
385   v[v>1]=1
386   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
387   out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
388   return(out)
389 }
390
391 # Función a maximizar
392 f_aux <- function(para){
393   D_u_0<-Du_Clay(u,v,para)-Du_Clay(u,vv,para)
394   D_u_0[D_u_0<=0]=1
395   D_u_0 <- log(D_u_0)
396   out<-(-sum(D_u_0))
397   return(out)
398 }
399
400 # Proceso de maximización
401 #para_ini <- theta2z(theta_ini ,family = family)
402 para_ifm <- optim(theta_ini ,f_aux,method = "BFGS").par
403 #theta_0 <- z2theta(para_ifm,family = family)
404 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
405 # para la regresión conjunta
406 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
407
408 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
409
410 # Creamos nuestra función a maximizar
411 #####
412 f_aux_reg <- function(para){
413
414   p <- ncol(var_x)
415   q <- ncol(var_x)
416   alpha <- para[1:p]
417   beta <- para[(p+1):(p+q)]
418   # theta <- z2theta(para[p+q+1],family )
419   theta <- z2theta(para[p+q+1],family = family)
420   delta <- para[p+q+2]
421   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
422   mu <- as.vector(exp(var_y%*%alpha))
423   #dummy <- density_joint(x,y,mu,delta ,lambda ,theta ,family=family ,zt=FALSE)
424   u2 <- pgam(data_model.Incurrido ,mu, delta )
425   v2<-ppois(y,lambda)
426   vv2<-ppois(y-1,lambda)
427
428   marginal.x <- dgam(data_model.Incurrido ,mu, delta )
429   marginal.x[marginal.x>=1]=0
430   marginal.x[marginal.x<=0]=0
431   marginal.x[is.na(marginal.x)]=0
432   marginal.x[marginal.x== -Inf]=0
433   marginal.x[marginal.x== Inf]=0
434
435   par_der <- Du_Clay(u2,v2,theta)
436   par_der1 <- Du_Clay(u2,vv2,theta)
437
438   dummy<- par_der-par_der1
439
440   dummy[y==0]=par_der[y==0]
441
442   out<-marginal.x*dummy
443   out[out<=0]=1e-10
444   ll <- -sum(log(out))

```

```

445      #if(negative==TRUE) ll <- (-ll)
446      return(ll)
447  }
448
449 ######
450 para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
451 a<-now()
452 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Estimación Maxima verosímil de los
453     parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
454 b<-now()
455 b-a
456 #load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
457     Parcial/JointModel_Gaus.RData")
458 para_optim.value # Log_Likelihood
459 vector <- para_optim.par
460 p       <- ncol(var_y)
461 q       <- ncol(var_y)
462 Log_Likelihood<-para_optim.value
463 # Parámetros optimizados
464 alpha_clay<-vector[1:p]
465 beta_clay<-vector[(p+1):(p+q)]
466 theta_clay=z2theta(vector[p+q+1],family) #-0.09506928
467 delta_clay<-vector[p+q+2]
468 tau_clay<-BiCopPar2Tau(par=theta_clay,family=family)#-0.06061453
469 head(beta_0)
470 head(beta_clay)
471 head(alpha_0)
472 head(alpha_clay)
473 # Determinamos los errores standar para el test de Wald
474 hessian_clay <- para_optim.hessian
475 Hinv          <- ginv(hessian_clay)
476 sd            <- sqrt(diag(Hinv))
477 sd.alpha_clay <-sd[1:p]
478 sd.beta_clay  <- sd[(p+1):(p+q)]
479 sd.theta_clay <-sd[p+q+1]
480
481 # Ordenamos los coeficientes estimados en una tabla
482 Reg_Cop_Clayton <- data.frame(Estimate_Frec=round(beta_clay,4),Std.error.freq=round(sd.beta_clay,4),z_
483     _value_frec=round(beta_clay/sd.beta_clay,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_clay/sd.beta_clay))),4),
484     Estimate_Sev=round(alpha_clay,4),Std.error.sev=round(sd.alpha_clay,4),z_value_sev=round(alpha_clay/sd.
485         alpha_clay,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_clay/sd.alpha_clay))),4))
486 #Reg_Cop_Clayton <- data.frame(Estimate_Frec=round(beta_clay,4),Std.error.freq=0,z_value_frec=0,"Prob_frec"
487     "=0",Estimate_Sev=round(alpha_clay,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
488 var_level<-row.names(Reg_Cop_Clayton)
489 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level
490     )
491
492 # Agregamos los interceptos
493 temp <- tibble(
494     var_level = c(
495         "Año2022",
496         "AS_Antiguedadb_1-2",
497         "AS_Canal_VentaCorredores",
498         "AS_Uso_VehParticular",
499         "AS_Marca_VehL2",
500         "AS_ZonaA3",
501         "AS_SAA.[1000-10000]"
502     ),
503     Estimate_Frec = 0
504     ,
505     Estimate_Sev = 0
506 )
507
508 # coeficientes del modelo de regresion con dependencia
509 Reg_Cop_Clayton <- Reg_Cop_Clayton %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_
510     Frec)`=round(exp(Estimate_Frec),4), `exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
511
512 Reg_Cop_Clayton <-Reg_Cop_Clayton %>% relocate(`exp(Estimate_Frec)`,.after=Prob_frec)
513
514 # Guardamos el modelo conjunto y los coeficientes
515 write.csv(Reg_Cop_Clayton, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
516     GLM_Conjuntos/Asistencia/Coef_Clay.csv", fileEncoding = "Latin1")
517 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
518     Conjuntos/Asistencia/JointModel_Clay.RData")
519
520 # Simulacion de la Pérdida Total

```

```

511 # =====
512 # CON DEPENDENCIA
513 # -----
514 lambda_loss_dep <- exp(var_x%*%beta_clay)*data_CM_Expuesto
515 mu_loss_dep <- exp(var_x%*%alpha_clay)
516 delta_loss_dep <- delta_clay
517 # Función de densidad fx
518 f_loss <-
519   function(loss ,mu,d, lam ,theta){
520     n<-length(loss)
521     if (length(lam)==1) lam <- rep(lam,n)
522     if (length(mu)==1) mu <- rep(mu,n)
523     out <- vector(length=n)
524
525     for(i in 1:n){
526       N <-1:20
527       v <-rpois(N, lam[i])
528       vv <-rpois(N-1, lam[i])
529       u <-pgam(loss[i]/N,mu[i],d)
530
531       Der_cop <-Du_Clay(u,v, theta )-Du_Clay(u,vv, theta )
532       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N
533       out[i] <-sum(dummy)
534     }
535
536     out[loss <=0]=0
537     return(out)
538   }
539
540 # Función de densidad acumulada Fx
541 F_loss <-
542   function(loss ,mu,d, lam ,theta){
543     out<-vector(length=length(loss))
544
545     for(i in 1:length(loss)){
546       floss <- function(s){
547         f_loss(s,mu[i],d, lam[i],theta )
548       }
549       out[i] <- integrate(floss ,0 ,loss[i]).value
550     }
551     return(out)
552   }
553
554
555 # Estimación del Total Loss (fuerte dedicacion computacional)
556
557 #k <-length(lambda_loss_dep)
558 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
559 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
560 L <- vector(length=k)
561 S <- vector(length=m)
562
563 for(j in 1:m){
564   r_uni<-runif(k)
565   for(i in 1:k){
566     f_root <-function(s){
567       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_clay)-r_uni[i]
568     }
569
570     tryCatch(
571       error = function(cnd) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
572       loss<-uniroot(f_root ,lower = 0,upper = 500000).root
573     )
574
575     #print(loss.root)
576     L[i]<-loss
577     perc<-paste(i/k*100,"%")
578     print(perc)
579   }
580   S[j]<-sum(L)
581   perc<-paste(j/m*100,"%")
582   print(perc)
583   #print(S[j])
584 }
585 S
586

```

```

587 # CON INDEPENDENCIA
588 # -----
589 lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
590     RIESGO COLECTIVO
591 mu_loss <-predict(glm.Sev.PP, newdata=data_CM, type='response')
592 delta_loss <-delta_0
593
594 M<-seq(1:10000)
595 for(i in 1:10000){
596   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
597   S<-vector(length=length(lambda_loss))
598   for(j in 1:length(muestra_cant)){
599     if(muestra_cant[j]==0){
600       S[j]=0
601     } else {
602       S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
603     }
604   }
605   M[i] <- sum(S)
606   porc<-paste((i/10000)*100,"%")
607   print(porc)
608   #print(M[i])
609 }
610
611 hist(M)
612 # Simulacion de la Frecuencia con Independencia
613 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
614 # Frecuencia
615 N<-seq(1:10000)
616 for(i in 1:10000){
617   N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response") .sim_1)/sum(data_model.Expuesto)
618   porc<-paste((i/10000)*100,"%")
619   print(porc)
620 }
621 hist(N)
622
623 #####
624 # Limpiamos el ambiente de variables
625 rm(list = ls())
626 gc()
627
628 #####
629 # REGRESION GUMBEL
630 #####
631 #####
632 #####
633 # Set Directory
634 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia")
635
636 # Carga los modelos GLM MARGINALES calculados en otros querys
637 load("glm.Frec.AS.RData")
638 load("glm.Sev.AS.RData")
639 load("Data_AS.RData")
640 load("Data_AS_model.RData")
641 load("Data_AS_model_CM.RData")
642
643 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal
644 # SEVERIDAD
645 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
646 var_x <- model.matrix(glm.Sev.AS) #data_CM[,3:12]
647 # FRECUENCIA
648 y <- data_model.Cantidad
649 Expuestos <- data_model.Expuesto
650 var_y <- model.matrix(glm.Frec.AS) #data[,3:12]
651
652 # Estimacion de parametros iniciales utilizando las distribuciones marginales
653 # SEVERIDAD
654 sd_alpha <- sqrt(diag(vcov(glm.Sev.AS)))
655 alpha_0 <- glm.Sev.AS.coefficients
656 delta_0 <- summary(glm.Sev.AS).dispersion
657 mu_0 <- exp(var_x%*%alpha_0)
658
659 # FRECUENCIA
660 beta_0 <- glm.Frec.AS.coefficients

```

```

661 sd.beta <- sqrt(diag(vcov(glm.Frec.AS)))
662 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
663
664 # REGRESION MEDIANTE COPULAS
665
666 # COPULA GUMBEL
667 family=4
668 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
669 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
  lambda_0)/(length(y)+1), family=family).par
670 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
  data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
671 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
672
673 # Estimación de las seudo-observaciones
674 u<-pgam(x,mu_0,delta_0)
675 v <- ppois(data_CM.Cantidad ,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
676 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
677
678 # Creamos la función primera derivada de la Función Copula
679 Du_Gum <- function(u,v,theta){
680   u[u<=0]=0.0001
681   u[u>1]=1
682   v[v<=0]=0.0001
683   v[v>1]=1
684   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
685   #out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
686   out <- (u^-1)*exp((-(-log(u))^theta+(-log(v))^theta)^(1/theta))
687   return(out)
688 }
689
690 # Función a maximizar
691 f_aux <- function(para){
692   D_u_0<-Du_Gum(u,v,para)-Du_Gum(u,vv,para)
693   D_u_0[D_u_0<=0]=1
694   D_u_0 <- log(D_u_0)
695   out<-(sum(D_u_0))
696   return(out)
697 }
698
699 # Proceso de maximización
700 #para_ini <- theta2z(theta_ini,family = family)
701 para.ifm <- optim(theta_ini,f_aux,method = "BFGS").par
702 #theta_0 <- z2theta(para.ifm,family = family)
703 theta_0 <- para.ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
  para la regresión conjunta
704 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
705
706 # Estimamos los coeficientes de regresión y el parámetro cópula mediante máxima verosimilitud
707
708 # Creamos nuestra función a maximizar
709 #####f_aux_reg <- function(para){
710 f_aux_reg <- function(para){
711
712   p <- ncol(var_x)
713   q <- ncol(var_x)
714   alpha <- para[1:p]
715   beta <- para[(p+1):(p+q)]
716   # theta <- z2theta(para[p+q+1],family)
717   theta <- z2theta(para[p+q+1],family = family)
718   delta <- para[p+q+2]
719   lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
720   mu <- as.vector(exp(var_y%*%alpha))
721   #dummy <- density_joint(x,y,mu,delta,lambda,theta,family=family,zt=FALSE)
722   u2 <- pgam(data_model.Incurrido ,mu, delta )
723   v2<-ppois(y,lambda)
724   vv2<-ppois(y-1,lambda)
725
726   marginal.x <- dgam(data_model.Incurrido ,mu, delta )
727   marginal.x[marginal.x>=1]=0
728   marginal.x[marginal.x<=0]=0
729   marginal.x[is.na(marginal.x)]=0
730   marginal.x[marginal.x== -Inf]=0
731   marginal.x[marginal.x== Inf]=0
732
733   par_der <- Du_Gum(u2,v2,theta)

```

```

734 par_der1 <- Du_Gum(u2 ,vv2 ,theta )
735
736 dummy<- par_der-par_der1
737
738 dummy[y==0]=par_der [y==0]
739
740 out<-marginal .x*dummy
741
742 out [ out <=0]=1e-10
743 11 <- -sum(log(out))
744
745 #if(negative==TRUE) 11 <- (-11)
746 return(11)
747 }
748
749 ##########
750 para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
751 a<-now()
752 para_optim <- optim(para_0,f_aux_reg ,method = "BFGS",hessian = TRUE) # Estimación Maxima verosmil de los
    parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
753 b<-now()
754 b-a #10.10871 mins
755 #load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
    Parcial/JointModel_Gaus.RData")
756 para_optim.value # Log_Likelihood
757 vector <- para_optim.par
758 p <- ncol(var_y)
759 q <- ncol(var_y)
760 Log_Likelihood<-para_optim.value
761 # Parámetros optimizados
762 alpha_gum<-vector[1:p]
763 beta_gum<-vector[(p+1):(p+q)]
764 theta_gum=z2theta(vector[p+q+1],family)
765 delta_gum<-vector[p+q+2]
766 tau_gum<-BiCopPar2Tau(par=theta_gum,family=family)
767 head(beta_0)
768 head(beta_gum)
769 head(alpha_0)
770 head(alpha_gum)
771
772 # Determinamos los errores estandar para el test de Wald
773 hessian_gum <- para_optim.hessian
774 Hinv <- ginv(hessian_gum)
775 sd <- sqrt(diag(Hinv))
776 sd.alpha_gum <-sd[1:p]
777 sd.beta_gum <- sd[(p+1):(p+q)]
778 sd.theta_gum <-sd[p+q+1]
779
780 # Ordenamos los coeficientes estimados en una tabla
781 Reg_Cop_Gumbel <- data.frame(Estimate_Frec=round(beta_gum,4),Std.error.freq=round(sd.beta_gum,4),z_value_
    freq=round(beta_gum/sd.beta_gum,4),"Prob_frec"=round(2*(1-pnorm(abs(beta_gum/sd.beta_gum))),4),Estimate_
    Sev=round(alpha_gum,4),Std.error.sev=round(sd.alpha_gum,4),z_value_sev=round(alpha_gum/sd.alpha_gum,4),"_
    Prob_Sev"=round(2*(1-pnorm(abs(alpha_gum/sd.alpha_gum))),4))
782 #Reg_Cop_Gumbel <- data.frame(Estimate_Frec=round(beta_gum,4),Std.error.freq=0,z_value_freq=0,"Prob_frec"
    ="0",Estimate_Sev=round(alpha_gum,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
783 var_level<-row.names(Reg_Cop_Gumbel)
784 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
785
786 # Agregamos los interceptos
787
788 temp <- tibble(
789   var_level = c(
790     "Año2022",
791     "AS_Antiguedadb_1-2",
792     "AS_Canal_VentaCorredores",
793     "AS_Uso_VehParticular",
794     "AS_Marca_VehL2",
795     "AS_ZonaA3",
796     "AS_SAA,[1000-10000]>"
797   ),
798   Estimate_Frec = 0
799   ,
800   Estimate_Sev = 0
801 )

```

```

804 # coeficientes del modelo de regresion con dependencia
805 Reg_Cop_Gumbel <- Reg_Cop_Gumbel %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)
806 `=round(exp(Estimate_Frec),4),`exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
807 Reg_Cop_Gumbel <-Reg_Cop_Gumbel %>% relocate(`exp(Estimate_Frec)` , .after=Prob_frec)
808
809 # Guardamos el modelo conjunto y los coeficientes
810 write.csv(Reg_Cop_Gumbel, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
811 GLM_Conjuntos/Asistencia/Coef_Gumb.csv", fileEncoding = "Latin1")
812 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_
813 Conjuntos/Asistencia/JointModel_Gumb.RData")
814
815 # Simulacion de la P閞dida Total
816 # =====
817 # CON DEPENDENCIA
818 # -----
819 lambda_loss_dep <- exp(var_x%*%beta_clay)*data_CM.Expuesto
820 mu_loss_dep <- exp(var_x%*%alpha_clay)
821 delta_loss_dep <- delta_clay
822
823 # Funci n de densidad fx
824
825 f_loss <-
826   function(loss, mu, d, lam, theta){
827     n<-length(loss)
828     if (length(lam)==1) lam <- rep(lam, n)
829     if (length(mu)==1) mu <- rep(mu, n)
830     out <- vector(length=n)
831
832     for(i in 1:n){
833       N <-1:20
834       v <-rpois(N, lam[i])
835       vv <-rpois(N-1, lam[i])
836       u <-pgam(loss[i]/N, mu[i], d)
837
838       Der_cop <-Du_Clay(u,v,theta)-Du_Clay(u,vv,theta)
839       dummy <-Der_cop*dgam(loss[i]/N, mu[i], d)/N
840       out[i] <-sum(dummy)
841     }
842
843     out[loss <=0]=0
844     return(out)
845   }
846
847 # Funci n de densidad acumulada Fx
848 F_loss <-
849   function(loss, mu, d, lam, theta){
850     out<-vector(length=length(loss))
851
852     for(i in 1:length(loss)){
853       floss <- function(s){
854         f_loss(s, mu[i], d, lam[i], theta)
855       }
856       out[i] <- integrate(floss, 0, loss[i]).value
857     }
858     return(out)
859   }
860
861 # Estimaci n del Total Loss (fuerte dedicacion computacional)
862
863 #k <-length(lambda_loss_dep)
864 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tama o de lambda)
865 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
866 L <- vector(length=k)
867 S <- vector(length=m)
868
869 for(j in 1:m){
870   r_uni<-runif(k)
871   for(i in 1:k){
872     f_root <-function(s){
873       F_loss(s, mu_loss_dep[i], delta_loss_dep, lambda_loss_dep[i], theta_clay)-r_uni[i]
874     }
875
876   tryCatch(

```

```

877     error = function(cnd) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
878     loss<-uniroot(f_root,lower = 0,upper = 500000).root
879   )
880
881   #print(loss.root)
882   L[i]<-loss
883   perc<-paste(i/k*100,"%")
884   print(perc)
885 }
886 S[j]<-sum(L)
887 perc<-paste(j/m*100,"%")
888 print(perc)
889 #print(S[j])
890 }
891 S
892
893 # CON INDEPENDENCIA
894 # -----
895 lambda_loss <-predict(glm.Frec.PP, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO
896 mu_loss <-predict(glm.Sev.PP, newdata=data_CM, type='response')
897 delta_loss <-delta_0
898
899 M<-seq(1:10000)
900 for(i in 1:10000){
901   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
902   S<-vector(length=length(lambda_loss))
903   for(j in 1:length(muestra_cant)){
904     if(muestra_cant[j]==0){
905       S[j]=0
906     } else{
907       S[j]=sum(rgam(n=muestra_cant[j],mu_loss[j],delta_loss ))
908     }
909   }
910   M[i] <- sum(S)
911   porc<-paste((i/10000)*100,"%")
912   print(porc)
913   #print(M[i])
914 }
915
916 hist(M)
917
918
919
920
921 # Simulacion de la Frecuencia con Independencia
922 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
923 # Frecuencia
924 N<-seq(1:10000)
925 for(i in 1:10000){
926   N[i] <- sum(simulate(glm.Frec.PP,nsim=1, type="response").sim_1)/sum(data_model.Expuesto)
927   porc<-paste((i/10000)*100,"%")
928   print(porc)
929 }
930 hist(N)
931 ######
932 # Limpiamos el ambiente de variables
933 rm(list = ls())
934 gc()
935
936 #####
937 # REGRESION FRANK
938 #####
939
940 # Set Directory
941 setwd("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/04_GLM_Marginales/Asistencia")
942
943 # Carga los modelos GLM MARGINALES calculados en otros querys
944 load("glm.Frec.AS.RData")
945 load("glm.Sev.AS.RData")
946 load("Data_AS.RData")
947 load("Data_AS_model.RData")
948 load("Data_AS_model_CM.RData")
949
950 # Creamos las variables de respuestas y la matriz de datos de cada glm marginal

```

```

951 # SEVERIDAD
952 x <- data_CM.Incurrido / data_CM.Cantidad #(Costo Medio)
953 var_x <- model.matrix(glm.Sev.AS) #data_CM[,3:12]
954 # FRECUENCIA
955 y <- data_model.Cantidad
956 Expuestos <- data_model.Expuesto
957 var_y <- model.matrix(glm.Frec.AS) #data[,3:12]
958
959 # Estimacion de parámetros iniciales utilizando las distribuciones marginales
960 # SEVERIDAD
961 sd_alpha <- sqrt(diag(vcov(glm.Sev.AS)))
962 alpha_0 <- glm.Sev.AS.coefficients
963 delta_0 <- summary(glm.Sev.AS).dispersion
964 mu_0 <- exp(var_x%*%alpha_0)
965
966 # FRECUENCIA
967 beta_0 <- glm.Frec.AS.coefficients
968 sd_beta <- sqrt(diag(vcov(glm.Frec.AS)))
969 lambda_0 <- exp(var_y%*%beta_0)*Expuestos
970
971 # REGRESION MEDIANTE COPULAS
972
973 # COPULA FRANK
974 family=5
975 # Estimación inicial del parámetro theta para estimarlo mediante el algoritmo de MPLE
976 theta_ini <- BiCopEst(rank(data_model.Incurrido-exp(var_y%*%alpha_0))/(length(y)+1),rank(y*Expuestos-
  lambda_0)/(length(y)+1), family=family).par
977 #theta_ini <- BiCopEst(rank(data_CM.Incurrido-exp(var_x%*%alpha_0))/(length(x)+1),rank(data_CM.Cantidad*-
  data_CM.Expuesto-exp(var_x%*%beta_0)*data_CM.Expuesto)/(length(x)+1), family=family).par
978 tau_ini <- BiCopPar2Tau(par=theta_ini,family = family)
979
980 # Estimación de las seudo-observaciones
981 u<-pgam(x,mu_0,delta_0)
982 v <- ppois(data_CM.Cantidad,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
983 vv <- ppois(data_CM.Cantidad-1,lambda = exp(var_x%*%beta_0)*data_CM.Expuesto)
984
985 # Creamos la función primera derivada de la Función Copula
986 Du_Frank <- function(u,v,theta){
987   u[u<=0]=0.0001
988   u[u>1]=1
989   v[v<=0]=0.0001
990   v[v>1]=1
991   #out <-pnorm((qnorm(v)-theta*qnorm(u))/(sqrt(1-theta^2)))
992   #out <- (u^(-theta)+v^(-theta)-1)^((-1/theta)-1)*(u^(-theta-1))
993   #out <- (u^(-1))*exp((-log(u))^theta+(-log(v))^theta^(1/theta))
994   out <- (exp(theta)*(exp(theta*v)-1))/(exp(theta*(u+1))+exp(theta*(v+1))-exp(theta)-exp(theta*(u+v)))
995   return(out)
996 }
997
998 # Función a maximizar
999 f_aux <- function(para){
1000   D_u_0<-Du_Frank(u,v,para)-Du_Frank(u,vv,para)
1001   D_u_0[D_u_0<=0]=1
1002   D_u_0 <- log(D_u_0)
1003   out<-(-sum(D_u_0))
1004   return(out)
1005 }
1006
1007 # Proceso de maximización
1008 #para_ini <- theta2z(theta_ini,family = family)
1009 para_ifm <- optim(theta_ini,f_aux,method = "BFGS").par
1010 #theta_0 <- z2theta(para_ifm,family = family)
1011 theta_0 <- para_ifm # Este será nuestro parámetro inicial estimado con el método IFM y que luego usaremos
  para la regresión conjunta
1012 tau <-BiCopPar2Tau(par=theta_0,family=family) # tau de Khendal para estimar el grado de dependencia
1013
1014 # Estimamos los coeficientes de regresión y el parámetro copula mediante máxima verosimilitud
1015
1016 # Creamos nuestra función a maximizar
1017 ##########
1018 f_aux_reg <- function(para){
1019
1020   p <- ncol(var_x)
1021   q <- ncol(var_y)
1022   alpha <- para[1:p]
1023   beta <- para[(p+1):(p+q)]
```

```

1024      # theta  <- z2theta(para[p+q+1],family)
1025      theta  <- z2theta(para[p+q+1],family = family)
1026      delta  <- para[p+q+2]
1027      lambda <- as.vector(exp(var_y%*%beta)*data_model.Expuesto)
1028      mu     <- as.vector(exp(var_y%*%alpha))
1029      #dummy <- density_joint(x,y,mu,delta,lambda,theta,family=family ,zt=FALSE)
1030      u2 <- pgam(data_model.Incurrido,mu,delta)
1031      v2<-ppois(y,lambda)
1032      vv2<-ppois(y-1,lambda)
1033
1034      marginal.x <- dgam(data_model.Incurrido ,mu, delta)
1035      marginal.x[marginal.x>=1]=0
1036      marginal.x[marginal.x<=0]=0
1037      marginal.x[is.na(marginal.x)]=0
1038      marginal.x[marginal.x==Inf]=0
1039      marginal.x[marginal.x== -Inf]=0
1040
1041      par_der   <- Du_Frank(u2,v2,theta)
1042      par_der1  <- Du_Frank(u2,vv2,theta)
1043
1044      dummy<- par_der-par_der1
1045
1046      dummy[y==0]=par_der[y==0]
1047
1048      out<-marginal.x*dummy
1049
1050      out[out<=0]=1e-10
1051      ll       <- -sum(log(out)) # Parámetros iniciales
1052
1053      #if(negative==TRUE) ll <- (-ll)
1054      return(ll)
1055  }
1056
1057 ##### Parámetros iniciales
1058 para_0<-c(alpha_0,beta_0,theta2z(theta_0,family=family),delta_0) # Parámetros iniciales
1059 a<-now()
1060 para_optim <- optim(para_0,f_aux_reg,method = "BFGS",hessian = TRUE) # Estimación Maxima verosmil de los
    parámetros (Toma hasta 3h sin matriz hessiana. 8h con Matriz hessiana)
1061 b<-now()
1062 b-a #5.13348 mins
1063 #load("C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_GLM_Conjuntos/Perdida_
    Parcial/JointModel_Gaus.RData")
1064 para_optim.value # Log_Likelihood
1065 para_optim.par
1066 vector <- para_optim.par
1067 p      <- ncol(var_y)
1068 q      <- ncol(var_y)
1069 Log_Likelihood<-para_optim.value # Parámetros optimizados
1070 alpha_frank<-vector[1:p]
1071 beta_frank<-vector[(p+1):(p+q)]
1072 theta_frank=z2theta(vector[p+q+1],family)
1073 delta_frank<-vector[p+q+2]
1074 tau_frank<-BiCopPar2Tau(par=theta_frank,family=family)
1075 head(beta_0)
1076 head(beta_frank)
1077 head(alpha_0)
1078 head(alpha_frank)
1079
1080 # Determinamos los errores estandar para el test de Wald
1081 hessian_frank <- para_optim.hessian
1082 Hinv        <- ginv(hessian_frank)
1083 sd          <- sqrt(diag(Hinv))
1084 sd.alpha_frank <-sd[1:p]
1085 sd.beta_frank <- sd[(p+1):(p+q)]
1086 sd.theta_frank <-sd[p+q+1]
1087
1088 # Ordenamos los coeficientes estimados en una tabla
1089 Reg_Cop_Frank <- data.frame(Estimate_Frec=round(beta_frank,4),Std.error.freq=round(sd.beta_frank,4),z_
    value_freq=round(beta_frank/sd.beta_frank,4),"Prob_freq"=round(2*(1-pnorm(abs(beta_frank/sd.beta_frank))),4),
    Estimate_Sev=round(alpha_frank,4),Std.error.sev=round(sd.alpha_frank,4),z_value_sev=round(alpha_frank/
    sd.alpha_frank,4),"Prob_Sev"=round(2*(1-pnorm(abs(alpha_frank/sd.alpha_frank))),4))
1090 #Reg_Cop_Frank <- data.frame(Estimate_Frec=round(beta_frank,4),Std.error.freq=0,z_value_freq=0,"Prob_freq"
    "=0,Estimate_Sev=round(alpha_frank,4),Std.error.sev=0,z_value_sev=0,"Prob_Sev"=0)
1091 var_level<-row.names(Reg_Cop_Frank)
1092 Reg_Cop_Frank <- Reg_Cop_Frank %>% as.tibble() %>% mutate(var_level=var_level) %>% relocate(var_level)
1093

```

```

1094 # Agregamos los interceptos
1095 temp <- tibble(
1096   var_level = c(
1097     "Año2022",
1098     "AS_AntiguedadB_1-2",
1099     "AS_Canal_VentaCorredores",
1100     "AS_Uso_VehParticular",
1101     "AS_Marca_VehL2",
1102     "AS_ZonaA3",
1103     "AS_SAa.[1000-10000]"
1104   ),
1105   Estimate_Frec = 0
1106 ,
1107   Estimate_Sev = 0
1108 )
1109
1110
1111 # coeficientes del modelo de regresión con dependencia
1112 Reg_Cop_Frank <- Reg_Cop_Frank %>% bind_rows(temp) %>% arrange(var_level) %>% mutate(`exp(Estimate_Frec)`=
1113   round(exp(Estimate_Frec),4), `exp(Estimate_Sev)`=round(exp(Estimate_Sev),4))
1114 Reg_Cop_Frank <-Reg_Cop_Frank %>% relocate(`exp(Estimate_Frec)`, .after=Prob_frec)
1115
1116 # Guardamos el modelo conjunto y los coeficientes
1117 write.csv(Reg_Cop_Frank, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/06_
1118 GLM_Conjuntos/Asistencia/Coef_Frank.csv", fileEncoding = "Latin1")
1119 save(para_optim, file="C:/Users/josephgarcia1/OneDrive - KPMG/Tesis_Actuarial/03_Modelos_GLM/
1120 Conjuntos/Asistencia/JointModel_Frank.RData")
1121
1122 # Simulación de la Pérdida Total
1123 # =====
1124
1125 # CON DEPENDENCIA
1126 # -----
1127 lambda_loss_dep <- exp(var_x%*%beta_frank)*data_CM.Expuesto
1128 mu_loss_dep <- exp(var_x%*%alpha_frank)
1129 delta_loss_dep <- delta_frank
1130
1131 # Función de densidad fx
1132 f_loss <-
1133   function(loss ,mu,d, lam ,theta){
1134     n<-length(loss)
1135     if (length(lam)==1) lam <- rep(lam,n)
1136     if (length(mu)==1) mu <- rep(mu,n)
1137     out <- vector(length=n)
1138
1139     for(i in 1:n){
1140       N <-1:20
1141       v <-ppois(N, lam[i])
1142       vv <-ppois(N-1, lam[i])
1143       u <-pgam(loss[i]/N,mu[i],d)
1144
1145       Der_cop <-Du_Frank(u,v,theta)-Du_Frank(u,vv,theta) # Primera derivada de la función copula
1146       dummy <-Der_cop*dgam(loss[i]/N,mu[i],d)/N # Teorela del Total Loss
1147       out[i] <-sum(dummy)
1148     }
1149
1150     out[loss <=0]=0
1151     return(out)
1152   }
1153
1154 # Función de densidad acumulada Fx
1155 F_loss <-
1156   function(loss ,mu,d, lam ,theta){
1157     out<-vector(length=length(loss))
1158
1159     for(i in 1:length(loss)){
1160       floss <- function(s){
1161         f_loss(s,mu[i],d, lam[i], theta)
1162       }
1163       out[i] <- integrate(floss ,0, loss[i]).value # Integral para el cálculo del a F(x) a partir de f(x)
1164     }
1165     return(out)
1166   }

```

```

1167 # Estimación del Total Loss (fuerte dedicacion computacional)
1168 #k <-length(lambda_loss_dep)
1169 k <- 100#length(lambda_loss_dep) # Cantidad de numeros aleatorios (Debe ser del tamaño de lambda)
1170 m <- 10 # Cantidad de muestras aleatorias (Deben ser 10 mil)
1171 L <- vector(length=k)
1172 S <- vector(length=m)
1173
1174
1175 for(j in 1:m){
1176   r_uni<-runif(k)
1177   for(i in 1:k){
1178     f_root <-function(s){
1179       F_loss(s,mu_loss_dep[i],delta_loss_dep,lambda_loss_dep[i],theta_frank)-r_uni[i]
1180     }
1181   }
1182
1183   tryCatch(
1184     error = function(cond) loss<-mu_loss_dep[i]*lambda_loss_dep[i],
1185     loss<-uniroot(f_root,lower = 0,upper = 500000).root
1186   )
1187
1188   #print(loss.root)
1189   L[i]<-loss
1190   perc<-paste(i/k*100,"%")
1191   print(perc)
1192 }
1193 S[j]<-sum(L)
1194 perc<-paste(j/m*100,"%")
1195 print(perc)
1196 #print(S[j])
1197 }
1198 S
1199
1200 # CON INDEPENDENCIA
1201 # -----
1202 lambda_loss <-predict(glm.Frec.RC, newdata=data_CM, type='response') # CAMBIAR LOS DATOS A UN MODELO DE
RIESGO COLECTIVO
1203 mu_loss <-predict(glm.Sev.RC, newdata=data_CM, type='response')
1204 delta_loss <-delta_
1205
1206 M<-seq(1:10000)
1207 for(i in 1:10000){
1208   muestra_cant<-rpois(length(lambda_loss),lambda_loss)
1209   S<-vector(length=length(lambda_loss))
1210   for(j in 1:length(muestra_cant)){
1211     if(muestra_cant[j]==0){
1212       S[j]=0
1213     } else{
1214       S[j]=sum(rgam(n =muestra_cant[j],mu_loss[j],delta_loss ))
1215     }
1216   }
1217
1218   M[i] <- sum(S)
1219   porc<-paste((i/10000)*100,"%")
1220   print(porc)
1221   #print(M[i])
1222 }
1223
1224 hist(M,breaks = 50)
1225 mean(M)
1226 sum(data_model.Incurrido)
1227 # Simulacion de la Frecuencia con Independencia
1228 #ADECUAR LOS DATOS DE SIMULACION A UN ESTRUCTURA DE MODELO COLECTIVO DE RIESGO
1229 # Frecuencia
1230 N<-seq(1:10000)
1231 for(i in 1:10000){
1232   N[i] <- sum(simulate(glm.Frec.RC,nsim=1, type="response").sim_1)/sum(data_model.Expuesto)
1233   porc<-paste((i/10000)*100,"%")
1234   print(porc)
1235 }
1236
1237 hist(N,breaks = 50)
1238 mean(N)
1239 sum(data_model.Cantidad)/sum(data_model.Expuesto)

```