ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

DESARROLLO DE UN MODELO DE PREDICCIÓN DE LA DESERCIÓN DE CLIENTES DE UNA EMPRESA DE MEDICINA PREPAGADA UTILIZANDO ANÁLISIS DE SUPERVIVENCIA

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

MARCOS AUGUSTO PÉREZ CÁRDENAS
marcos.perez.c@outlook.com

DIRECTOR: ING. MIGUEL FLORES SÁNCHEZ, PHD.
miguel.flores@epn.edu.ec

Quito, septiembre 2020

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Marcos Augusto Pérez Cárdenas, bajo mi supervisión.

MIGUEL ALFONSO FLORES SANCHEZ

Firmado digitalmente por MIGUEL ALFONSO FLORES SANCHEZ
Fecha: 2020.09.16 09:13:22 -05'00'

Ing. Miguel Flores Sánchez, PhD.

DIRECTOR DE PROYECTO

DECLARACIÓN

Yo, Marcos Augusto Pérez Cárdenas , declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Marcos Augusto Pérez Cárdenas

DEDICATORIA

A mis hijas, Itzel y Zoe, que han sido mi motorsito durante estos años, y a mi esposa Leonella, que siempre ha estado ahí para apoyarme y a veces regañarme, a las tres las amo mucho.

A mis padres, Lourdes y Arturo, que me educaron y me aman como yo los amo, a mis hermanos, Arturo y Marcelo, con quienes crecí y compartí tantas hermosas experiencias, siempre están en mi corazón.

AGRADECIMIENTOS

A mis padres, Lourdes y Arturo, por todo su apoyo durante mis estudios, por su amor incondicional, los amo mucho.

A mi esposa Leonella por apoyarme y animarme todo el tiempo.

A mi director de tesis Miguel, por su paciencia, guía y consejos.

A Humana S.A., una empresa comprometida con la salud, la educación y el desarrollo del país y también a todos quienes hicieron posible este trabajo con su apoyo y consejos.

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE CÓDIGOS FUENTE

RESUMEN

Las empresas dan hoy más importancia a la retención de clientes. Atraer a un consumidor nuevo puede llegar a costar hasta cinco veces lo que cuesta retener a un cliente existente. Económicamente es mucho mejor retener los clientes existentes que atraer nuevos clientes y, además, genera mucho más valor a la empresa, no solamente el beneficio económico. El presente trabajo está enfocado en predecir, mediante el Modelo de Cox, la deserción de clientes de una empresa de Medicina Prepagada.

Este modelo se desarrolló mediante el software R, usando principalmente el paquete "survival" y de manera secundaria el paquete "survminer". El paquete "survival" permite no solamente el desarrollo en si del modelo, sino también la validación de los supuestos del modelo de Cox y la evaluación de la eficacia del modelo. El paquete "survminer" nos permite principalmente la visualización de los resultados (curvas de supervivencia, diagramas de censura, entre otros) de una manera más clara y además contiene otras funciones disponibles para trazar curvas ajustadas para el modelo Cox y examinar visualmente los supuestos del mismo.

Algunos de los resultados obtenidos coinciden con estudios realizados en otros países sobre la deserción de clientes, no precisamente en el mismo sector de la medicina prepagada. Se podrá ver, además, que las dificultades para predecir el comportamiento de los clientes son las mismas en una industria y otra, guardando por supuesto las particularidades de cada una. Los modelos obtenidos fueron validados con una base de prueba y cada uno tuvo particulares hallazgos.

Palabras clave: análisis de supervivencia, modelo de Cox, deserción, modelo de riego proporcionales.

ABSTRACT

Companies today place more importance on customer retention. Attracting a new consumer can cost up to five times what it costs to retain an existing customer. Economically it is much better to retain customers who can attract new customers and, in addition, it generates much more value to the company, not only the economic benefit. The present work is focused on predicting, through the Cox Model, the desertion of clients of a Prepaid Medicine company.

This model is used through the R software, mainly using the “survival” package and secondarily the “survminer” package. The “survival” package allows not only the development of the model itself, but also the validation of the assumptions of the Cox model and the evaluation of the effectiveness of the model. The “survminer” package mainly allows us to visualize the results (survival curves, censorship diagrams, among others) in a clearer way and also contains other functions available to plot fitted curves for the Cox model and visually examine the assumptions of the same.

Some of the results obtained coincide with studies carried out in other countries on customer desertion, not exactly in the same sector of prepaid medicine. It will also be seen that the difficulties in predicting the behavior of customers are the same in one industry and another, keeping of course the particularities of each one. The models obtained were validated with a test base and each one had its own particular findings.

Key words: survival analysis, Cox model, churn, proportional hazard model.

1 INTRODUCCIÓN

El área de mercadeo, con el fin de apuntalar el cumplimiento de los objetivos estratégicos de las compañías, invierte fuertemente en el control de la deserción de clientes ya que ésta, a más de impedir que se logren los objetivos de participación y penetración del mercado, se traduce inmediatamente, por obvias razones, en perdida de los ingresos que, de pasar ciertos límites y dada la dinámica de competencia de los mercados, cuestan mucho recuperar. Las estrategias de marketing que se usan para mantener clientes, además de encaminarse a garantizar penetración de mercado, se convierten también en esquemas para generar una sólida relación cliente - empresa que desemboca en última instancia en fidelización del cliente. Clientes fieles generan un valioso beneficio para las empresas, el aseguramiento de los ingresos. Cuando las tasas de fidelización de clientes son altas, los indicadores de valor de marca y reputación, aumentan en niveles nada despreciables, generando valor a la marca [1]. De esta manera, un adecuado control de la deserción redunda de manera inequívoca en el aumento de los índices de lealtad, de fidelización de clientes y valor de marca.

Las empresas dan hoy más importancia a la retención de clientes. Atraer a un consumidor nuevo puede llegar a costar hasta cinco veces lo que cuesta retener a un cliente existente [2]. Económicamente es mucho mejor retener los clientes existentes que atraer nuevos clientes y, además, como vimos antes, genera mucho más valor a la empresa, no solamente el beneficio económico.

A continuación, veremos algunos cálculos interesantes sobre la retención de clientes. Estos datos son presentados por Kotler y Keller en su libro Dirección de Marketing[2] en la página 156[1], pero a su vez fueron tomados del libro “The Loyalty Effect”[3] de Frederick F. Reichheld:

[1] Se toma en cuenta la versión traducida al español

1. Adquirir nuevos clientes cuesta cinco veces más que satisfacer y retener a los clientes existentes. Para conseguir que un cliente satisfecho abandone a su proveedor actual es necesario hacer muchos esfuerzos.
2. La empresa promedio pierde el 10 % de sus clientes al año.
3. Reducir un 5 % el índice de abandono de clientes puede aumentar las utilidades entre un 25 % y 85 %, en función del sector de que se trate.
4. El índice de ganancias por cliente tiende a aumentar con el tiempo, siempre que el cliente se mantenga como tal.

Aunque los cálculos originales fueron hechos para la industria de las tarjetas de crédito principalmente, nos muestran el altísimo impacto que puede tener mejorar la retención de clientes.

En general, la pérdida de clientes es un mal que afecta a todas las empresas sin excepción y en particular las empresas de medicina prepagada sufren una doble afectación debida a este problema, por un lado, pierden los ingresos que estos clientes generaban a la empresa y por otro, el perder volumen de clientes y primas hace necesario que los precios de sus productos suban para poder controlar el riesgo de pérdidas económicas, esto último no es tan intuitivo pero se pretende demostrarlo dentro del ámbito del trabajo propuesto.

En pocas palabras, mientras más clientes tenga una empresa mayor será su capacidad de soportar eventualidades que puedan causar problemas financieros a la empresa.

Por las razones antes expuestas, las empresas de medicina prepagada no pueden perder muchos clientes. La compañía que auspicia esta investigación tiene actualmente una deserción cercana al 2,5 % mensual, lo que quiere decir que en un año la empresa pierde al menos un 30 % de sus clientes los cuales son reemplazados por clientes nuevos, no obstante, el nivel de deserción es demasiado alto (Como ya vimos una empresa promedio pierde solo un 10 % de sus clientes al año) y se necesita saber cómo evitar que más clientes sigan terminando su relación contractual con la compañía.

El proceso para tratar de mitigar este problema es meramente reactivo y poco eficiente. Los asesores de servicio al cliente reciben una carta del cliente en la que el contratante indica su deseo de desafiliarse, es decir, dar por terminado el contrato con la empresa. Como respuesta a esto, los asesores suelen llamar al cliente y tratan de convencerlo de quedarse, le ofrecen descuentos y promociones, pero la mayoría de veces no funciona.

Las empresas suelen tener una base de datos donde se registra los motivos por los cuales los clientes tomaron la decisión de desafiliarse y la empresa auspiciante no es la excepción, sin embargo, esta medida es insuficiente y además muy poco efectiva ya que es muy común que los clientes no revelen los motivos reales de su decisión con el fin de evitar el engorroso proceso de desafiliación lo más que puedan y la misma llamada de los asesores de servicio al cliente suele acabar por molestarlos aún más. En otras ocasiones los clientes suelen colaborar mucho pero no revelan a profundidad sus motivos reales. Todo lo anterior hace que pocas veces se tenga el motivo real y a detalle de la razón de su desafiliación, de hecho, se cree que menos del 80 % de los datos recopilados de las desafiliaciones son verídicos.

Por todo lo antes expuesto, se vuelven necesarias dos cosas, la primera es que se debe identificar los motivos principales por los que los clientes deciden desafiliarse y lo segundo es que se hace necesario un modelo predictivo que permita anticiparse a que el cliente decida desafiliarse y tomar las medidas necesarias para retenerlo al menos en la mayoría de los casos.

Un modelo predictivo no solo ayudaría a anticiparse a la deserción, sino que también revelaría los motivos verdaderos por los que los clientes se desafilian y pesaría los motivos unos contra otros, lo que ayudaría a atacar esos motivos y mejorar los procesos y servicios que están causando que los clientes tomen la decisión de terminar la relación con la empresa.

El presente trabajo se enfoca en tratar de predecir la deserción del cliente mirando las variables de servicio, variables relacionadas a la venta y también a la naturaleza del producto, pero ¿qué nos asegura que estas variables tendrán alguna relación con la deserción y que por lo tanto se puede hacer un modelo predictivo?

En un estudio hecho para la industria optométrica en Sudáfrica se utilizó el análisis de regresión múltiple para explicar la relación entre las variables independientes de satisfacción del cliente, confianza y compromiso, y la variable dependiente lealtad del cliente. Los resultados mostraron que las tres variables independientes, a saber, la satisfacción del cliente, la confianza y el compromiso, tuvieron una influencia positiva significativa en la variable dependiente de la lealtad del cliente. Además, El coeficiente de correlación de Pearson también se calculó para investigar más a fondo la correlación entre las variables independientes de la satisfacción del cliente, la confianza y el compromiso, y la variable dependiente de la lealtad del cliente. Los hallazgos confirmaron que todas las variables independientes tuvieron una influencia significativa en la variable dependiente de la lealtad del cliente. La satisfacción

del cliente tuvo la mayor influencia en la lealtad del cliente en comparación con la confianza y el compromiso (R^2 = 0,732; p-valor = 0,855). La confianza ocupó el segundo lugar en términos de su influencia en la lealtad del cliente, en comparación con la satisfacción del cliente y el compromiso (R^2 = 0,703; p-valor = 0,835). El compromiso tuvo la tercera mayor influencia en la lealtad del cliente después de la satisfacción del cliente y la confianza (R^2 = 0,53; p-valor = 0,728) [4].

No obstante, si bien la satisfacción del cliente es muy influyente en la lealtad del cliente también se ha encontrado que el 85 % de los pacientes de la industria de la óptica que desertan a otros proveedores de servicios se mostraron satisfechos con el servicio recibido justo antes de desertar [5]. Esto quiere decir que dar un buen servicio no es suficiente para evitar la deserción de clientes.

Por otra parte, también es importante saber con qué tipo de cliente se debe trabajar para elaborar el modelo; una mala elección del cliente objetivo nos podría llevar a conclusiones erradas o resultados inútiles.

La empresa patrocinadora se enfoca en 5 líneas de negocio, dos de las cuales, debido a la naturaleza del producto, tienen poca frecuencia de uso, es decir, los clientes suelen usar muy poco el servicio y parte del objetivo de este trabajo es saber cómo mejorar ese servicio, de manera que estas dos líneas quedarían descartadas del análisis debido a que los resultados no ayudarían a despejar dudas importantes.

Las tres líneas de negocio restantes son:

- **Corporativa**. Se refriere a clientes que, por lo general, son empresas grandes o grupos comerciales. Estas empresas buscan dar un beneficio adicional a sus empleados con un plan de medicina prepagada. La empresa patrocinadora diseña planes a la medida de las necesidades de estas empresas y ofrece productos específicos para cada cliente, de manera que el cliente tiene varios productos personalizados para cubrir las necesidades de sus empleados, en función de lo que ellos consideren adecuado.

- **Individual**. Para este caso los clientes son personas independientes y sus familias. Estas personas tienen acceso a planes prediseñados en función del mercado y de las necesidades generales que tiene la población ecuatoriana. Estos planes son certificados por los organismos competentes, quienes dan la aprobación para que sean comercializados luego de verificar que cada producto cumpla con los estándares ne-

cesarios.

- **Empresarial**. Es una suerte de combinación de las dos líneas de negocio anteriores. Al igual que en la línea de negocio individual se tienen planes prediseñados pero los clientes son asociaciones, organismos o empresas pequeñas que deben contar con algunas condiciones mínimas para poder acceder a estos planes. Los clientes negocian los planes para cada uno de sus propios socios, afiliados o empleados de manera que obtienen beneficios para todos. Estos planes son certificados por los organismos competentes, quienes dan la aprobación para que sean comercializados luego de verificar que cada producto cumpla con los estándares necesarios.

Trabajar en este proyecto con clientes de la línea de negocio Corporativa ocasionaría muchas desventajas debido a que tiene una variabilidad enorme de planes, ya que cada cliente es diferente y solicita diferentes beneficios para sus empleados, por ejemplo, en las empresas que tienen obreros en zonas de difícil acceso solicitan ambulancia aérea como requisito indispensable para sus planes mientras empresas cuyos empleados están básicamente en oficina ni siquiera lo consideran. Sumado a lo anterior está el hecho de que cada empresa o corporación tienen desde planes para obreros, cuyos montos de cobertura suelen ser muy modestos, hasta planes para gerentes, los cuales tienen todo tipo de beneficios y montos de cobertura altos. Además, el comportamiento de sus empleados respecto al plan de medicina prepagada suele ser muy heterogéneo ya que los obreros suelen usar su plan para casos muy puntuales y la mayoría de las veces prefieren usar el servicio del Seguro Social que es gratuito y no tienen que pagar copago, por muy bajo que este sea, mientras que gerentes y jefes hacen uso del plan con muchísima frecuencia.

El comportamiento de los clientes de la línea de negocio Empresarial suele ser mucho más estable pero la cantidad de clientes es muy baja y sumado a esto los contratos, si bien tienen planes prediseñados, suelen tener clausulas específicas de negociación con el cliente. Además, es difícil identificar la deserción ya que un socio de una organización, por ejemplo, podría desafiliarse del plan de medicina prepagada sin que esto implique necesariamente que el contrato se cancele, si aún quedan cinco socios o más el contrato seguiría vigente y esto impediría que se pueda identificar algún tipo de descontento con el servicio o la intención de desafiliarse de la mayoría.

Por lo expuesto anteriormente es preferible trabajar únicamente con los contratos de la línea de negocio Individual que tiene una buena cantidad de clientes, el comportamiento de los

clientes es bastante homogéneo y la deserción se puede medir de manera efectiva.

Para el presente trabajo se seleccionó como cliente objetivo a todos los clientes de la Línea de Negocio Individual.

1.1 OBJETIVO GENERAL

Desarrollar un análisis de supervivencia para la predicción de deserción de clientes de una empresa de medicina prepagada de manera que permita prevenir el riesgo de caída de cartera y determinar el perfil de los clientes.

1.2 OBJETIVOS ESPECÍFICOS

1. Evaluar los riesgos actuales de la empresa con respecto algunos escenarios de la deserción de clientes.
2. Determinar el perfil de los clientes que desertan con la finalidad de que se propongan estrategias de retención.
3. Desarrollar e implementar una función en lenguaje R que permita automatizar el análisis de supervivencia.

1.3 HIPÓTESIS

Se puede predecir la deserción de un cliente a partir de eventos específicos donde el cliente tuvo una mala experiencia relacionada con el servicio.

2 ANÁLISIS DE SUPERVIVENCIA Y EL MODELO DE COX

Primeramente, se presentará el análisis de supervivencia con las definiciones y los resultados que apalancan el modelo de riesgos proporcionales como una poderosa herramienta. Luego, se explicará el modelo de Cox de riesgos proporcionales, empezando por una breve explicación de los modelos tradicionales, por así decirlo, que han sido usados para la predicción de deserción de clientes (El modelo logit, las redes neuronales y los árboles de decisión) de manera de dar al lector todo un abanico de posibles maneras de solucionar el problema objetivo de este trabajo. En la justificación metodológica veremos a profundidad por qué es preferible el uso del modelo de riesgos proporcionales para el problema de caída de cartera, por encima de otros modelos.

2.1 ANÁLISIS DE SUPERVIVENCIA

Klein y Moeschberger definen el análisis de supervivencia como una “Técnica que permite describir el comportamiento de los datos que corresponden a tiempo o duración desde un origen hasta la ocurrencia de un cambio”[6].

El análisis de supervivencia requiere una variable dependiente que representa el periodo de tiempo entre un suceso inicial y uno final. El análisis de supervivencia nos permite estimar el efecto que una o más variables independientes tienen sobre la variable dependiente y por extensión sobre el evento final. De esta manera podemos saber cuáles son las posibles causas de la ocurrencia del evento final o al menos que hecho es el que más incide sobre éste.

Es decir, dada una variable que representa el tiempo que transcurre hasta que se produce un determinado evento final, el análisis al que nos referimos permite estimar, en función del tiempo, la probabilidad de que ocurra este evento.

En conclusión, El análisis de supervivencia es una técnica derivada del control estadístico

de calidad que nos permite predecir el tiempo que tardará en ocurrir el evento de interés y las variables que inciden en la ocurrencia de este hecho.

En términos de estadística podríamos decir que se puede entender como la probabilidad de que ocurra el evento de interés en un periodo p, dado que no ocurrió en un periodo p-1.

Este análisis permite usar variables continuas, variables ordinales, variables categóricas y estas pueden representar todo tipo de factores de influencia sobre el evento final, por ejemplo, pueden representar magnitudes, medidas, indicadores, variables geográficas, edad, genero, y cada variable puede verse en diferentes periodos temporales. Esto hace que el modelo cuente con una enorme versatilidad.

Para este análisis debemos tener claro algunos conceptos que se explican a continuación.

2.1.1 LA FUNCIÓN DE SUPERVIVENCIA

Cuantifica la probabilidad de que un individuo sobreviva (no le ocurra el evento de interés) más allá del tiempo t. Más formalmente podemos presentarla como: Sea T una variable aleatoria continua y positiva cuya función de distribución es $F(t)$ y cuya función de densidad es $f(t)$ entonces su correspondiente función de supervivencia $S(t)$ se define como

$$S(t) = 1 - F(t) = \mathcal{P}(T > t) \tag{2.1}$$

Es importante tener en cuenta que la función de supervivencia es una función no creciente con un valor de 1 en el origen y 0 cuando t tiende al infinito. Si T es una variable aleatoria continua, entonces, $S(t)$ es una función continua, estrictamente decreciente.

Cuando T es una variable aleatoria continua, la función de supervivencia es el complemento de la función de distribución acumulativa, es decir, $S(t) = 1 - F(x)$, donde $F(t) = \mathcal{P}(T \leq t)$. Además, la función de supervivencia es la integral de la función de densidad de probabilidad, $f(t)$, es decir,

$$S(t) = \mathcal{P}(T > t) = \int_t^\infty f(x)dx \tag{2.2}$$

De esta manera,

$$f(t) = -\frac{dS(t)}{dt}$$

Nótese que $f(t)\Delta t$ se puede considerar como la probabilidad “aproximada” de que el evento

de interés ocurra en el momento t y que $f(t)$ es una función no negativa con el área bajo la curva igual a uno.

Cuando T es una variable aleatoria discreta, se requieren diferentes técnicas. Las variables aleatorias discretas en los análisis de supervivencia surgen debido al redondeo de las mediciones, la agrupación de los tiempos de falla en intervalos, o cuando las vidas se refieren a un número integral de unidades. Supongamos que T puede tomar valores $t_j, j = 1, 2, ...,$ con función de densidad de probabilidad $p(t_j) = \mathcal{P}(T = t_j), j = 1, 2, ...,$ donde $t_1 < t_2 < \cdots$. La función de supervivencia para una variable aleatoria discreta T viene dada por

$$S(t) = \mathcal{P}(T > t) = \sum_{t_j > t} p(t_j) \tag{2.3}$$

2.1.2 LA FUNCIÓN DE RIESGO

Esta función es fundamental en el análisis de riesgo. Esta función también se conoce como la tasa de falla instantánea, la fuerza de mortalidad en la demografía, la función de intensidad en procesos estocásticos, la tasa de falla específica por edad en epidemiología, la inversa de la relación de Mill en economía, o simplemente como la razón de riesgo.Puede verse como la probabilidad de que a un individuo le ocurra el evento de interés en la siguiente unidad de tiempo Δt dado que ha sobrevivido hasta el tiempo t. Más formalmente se define como

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \tag{2.4}$$

Si T es una variable aleatoria continua, entonces,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \ln[S(t)]}{dt} \tag{2.5}$$

La función de riesgo acumulada $H(t)$, se define por

$$H(t) = \int_0^t h(u)du = -\ln[S(t)] \tag{2.6}$$

Por lo tanto, para líneas de tiempo continuas,

$$S(t) = exp[-H(t)] = exp\left[-\int_0^t h(u)du\right] \tag{2.7}$$

A partir de 2.4, se puede ver que $h(t)\Delta t$ puede verse como la probabilidad "aproximada" de que un individuo de edad t experimente el evento en el siguiente instante. Esta función es particularmente útil para determinar las distribuciones de falla apropiadas utilizando información cualitativa sobre el mecanismo de falla y para describir la forma en que la probabilidad de experimentar el evento cambia con el tiempo. La única restricción para $h(t)$ es que debe ser no negativa, es decir, $h(t) \geq 0$.

Cuando T es una variable aleatoria discreta, la función de riesgo viene dada por

$$h(t_j) = \mathcal{P}(T = t_j | T \geq t_j) = \frac{p(t_j)}{S(t_{j-1})}, \quad j = 1, 2, ... \tag{2.8}$$

donde $S(t_0) = 1$, debido a que al reemplazar $p(t_j) = S(t_{j-1}) - S(t_j)$ en la ecuación 2.8 tenemos que $h(t_j) = 1 - S(t_j)/S(t_{j-1}),\ j = 1, 2, \cdots$.

Debemos tener en cuenta que la función de supervivencia puede escribirse como el producto de probabilidades de supervivencia condicional

$$S(t) = \prod_{t_j \leq t} S(t_j)/S(t_{j-1}) \tag{2.9}$$

Luego entonces, la función de supervivencia se relaciona con la función de riesgo por

$$S(t) = \prod_{t_j \leq t} [1 - h(t_j)] \tag{2.10}$$

2.1.3 LA FUNCIÓN VIDA RESIDUAL MEDIA Y LA MEDIA DE VIDA

El siguiente parámetro básico de interés en los análisis de supervivencia es la vida residual media (o mrl por su siglas en inglés) en el tiempo t. Para los sujetos de edad t, este parámetro mide su vida útil restante esperada. Se define como $mrl(t) = E(T - t | T > t)$. Se puede demostrar que la vida residual media es el área bajo la curva de supervivencia a la derecha de t dividida por $S(x)$. Nótese que la media de vida, $\boldsymbol{\mu} = mrl(0)$, es el área total bajo la curva de supervivencia.

Para una variable aleatoria continua,

$$mrl(t) = \frac{\int_t^\infty (x - t) f(x) dx}{S(t)} = \frac{\int_t^\infty S(x) dx}{S(t)} \tag{2.11}$$

y

$$\boldsymbol{\mu} = E(T) = \int_0^\infty xf(x)dx = \int_0^\infty S(x)dx \tag{2.12}$$

Además, la varianza de T está relacionada con la función de supervivencia por

$$Var(T) = 2\int_0^\infty xS(x)dx - \left[\int_0^\infty S(x)dx\right]^2 \tag{2.13}$$

El p-ésimo cuantil de la distribución de T es t_p tal que

$$F(t_p) \geq p \quad y \quad S(t_p) \geq 1 - p \tag{2.14}$$

Si T es una variable aleatoria continua, entonces el p-ésimo cuantil se encuentra resolviendo la ecuación $S(t_p) = 1 - p$. Se sigue que la mediana de vida de una variable aleatoria continua T es el valor $x_{1/2}$ tal que $S(x_{1/2}) = 1/2$.

2.1.4 ESTIMADOR DE KAPLAN Y MEIER

El estimador estándar de la función de supervivencia, propuesto por Kaplan y Meier [7], se llama estimador Producto-Límite. Se trata de un estadístico no paramétrico que se utiliza para estimar la ya mencionada función a partir de datos de tiempos de supervivencia de los individuos.

El estimador de la función de supervivencia $S(t)$ (la probabilidad de que la vida sea más larga que t) viene dado por:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right) \tag{2.15}$$

Donde t_i es el tiempo en el que ocurrió al menos un evento, d_i el número de eventos que sucedieron hasta el instante t_i y Y_i los individuos en riesgo (es decir, que aún no les ha ocurrido el evento de interés o son datos censurado) hasta t_i, para $i = 1, 2, ..., k$ con k $(k \leq n)$ el número de individuos al que les ocurrió el evento de interés en una población de tamaño n.

El estimador Producto-Límite es una función escalonada con saltos en los tiempos de eventos observados. El tamaño de estos saltos depende no solo del número de eventos observados en cada momento del evento t_i, sino también del patrón de las observaciones censuradas antes de t_i.

La varianza del estimador se puede estimar mediante la fórmula de Greenwood:

$$\widehat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

El error estándar del estimador de Producto-Límite viene dado por $\{\widehat{Var}[\hat{S}(t)]\}^{1/2}$.

Una acotación más, para valores de t más allá del mayor tiempo de observación, este estimador no está bien definido.

2.1.5 LA PRUEBA LOG-RANK

La prueba de log-rank es una prueba de hipótesis para comparar las tasas de riesgo de K $(K \geq 2)$ poblaciones. Es una prueba no paramétrica y apropiada para usar cuando existen datos censurados. Aunque fue propuesta inicialmente por Mantel en 1966[8], con el tiempo varios autores han ido extendiendo el concepto original hasta tener una herramienta con múltiples opciones y muy versátil.

Según Klein y Moeschberger [6] explican que con esta prueba pretendemos probar el siguiente conjunto de hipótesis:

$$H_0 : \ h_1(t) = h_2(t) = \cdots = h_K(t), \ para\ todo\ t \leq \tau \tag{2.16}$$

contra la hipótesis alternativa

$$H_A : \ Al\ menos\ uno\ de\ los\ h_j(t)\ con\ j = 1, ..., K\ es\ diferente\ para\ algún\ t \leq \tau.$$

La hipótesis alternativa es global en el sentido de que deseamos rechazar la hipótesis nula si, al menos, una de las poblaciones difiere de las demás en algún momento. También se presentará pruebas que son más poderosas en el caso de alternativas ordenadas.

Para probar la hipótesis 2.16 podemos tener muestras independientes con datos censurados y, posiblemente, truncadas a la izquierda para cada una de las K poblaciones. Sean $t_1 < t_2 < \cdots < t_D$ los distintos tiempos de muerte (o tiempo en el que ocurre el evento de interés) en la muestra combinada. En el momento t_i observamos d_{ij} eventos en la j-ésima muestra sin contar los Y_{ij} individuos en riesgo, $j = 1, ..., K$, $i = 1, ..., D$. Sea $d_i = \sum_{j=1}^{K} d_{ij}$ y $Y_i = \sum_{j=1}^{K} Y_{ij}$ el número de muertes y el número de individuos en riesgo en la muestra

combinada en el tiempo t_i, $i = 1, ..., D$.

Para Klein y Moeschberger [6] la prueba de H_0 se basa en comparaciones ponderadas de la tasa de riesgo estimada de la j-ésima población bajo las hipótesis nula y alternativa, basadas en el estimador Nelson-Aalen. Si la hipótesis nula es cierta, entonces, un estimador de la tasa de riesgo esperada en la j-ésima población bajo H0 es el estimador de muestra agrupada de la tasa de riesgo d_i/Y_i. Utilizando solo datos de la j-ésima muestra, el estimador de la tasa de riesgo es d_{ij}/Y_{ij}. Para hacer comparaciones, sea $W_j(t)$ una función de ponderación positiva con la propiedad de que $W_j(t_i)$ es cero siempre que Y_ij es cero. La prueba de H_0 se basa en el estadístico:

$$Z_i(\tau) = \sum_{i=1}^{D} W_j(t_i) \left\{ \frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_i} \right\}, \quad j = 1, ..., K \tag{2.17}$$

Si todas las $Z_j(\tau)$ son cercanas a cero, entonces, hay poca evidencia para creer que la hipótesis nula es falsa, mientras que, si una de las $Z_j(\tau)$ está lejos de cero, entonces, hay evidencia de que esta población tiene una tasa de riesgo diferente de la esperada bajo la hipótesis nula.

Klein y Moeschberger [6] afirman que aunque la teoría general permite diferentes funciones de ponderación para cada una de las comparaciones en 2.17, en la práctica, todas las pruebas que se utilizan comúnmente tienen una función de ponderación $W_j(t_i) = Y_{ij}W(t_i)$. Aquí, $W(t_i)$ es un peso común compartido por cada grupo, y Y_{ij} es el número en riesgo en el j-ésimo grupo en el tiempo t_i. Con esta elección de funciones de peso

$$Z_i(\tau) = \sum_{i=1}^{D} W(t_i) \left\{ d_{ij} - Y_{ij} \frac{d_i}{Y_i} \right\}, \quad j = 1, ..., K \tag{2.18}$$

Debemos tener en cuenta que con esta clase de ponderaciones, el estadístico de prueba es la suma de la diferencia ponderada entre el número observado de muertes y el número esperado de muertes bajo H_0 en la j-ésima muestra. El número esperado de muertes en la muestra j al tiempo t_i es la proporción de individuos en riesgo Y_{ij}/Y_i que están en la muestra j en el momento t_i, multiplicado por el número de muertes en el momento t_i.

La varianza de $Z_j(\tau)$ en 2.18 está dada por

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i} \left(1 - \frac{Y_{ij}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i, \quad j = 1, ..., K \tag{2.19}$$

y la covarianza de $Z_j(\tau)$, $Z_g(\tau)$ se expresa por

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i} \frac{Y_{ig}}{Y_i} \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i, \quad g \neq j \tag{2.20}$$

El término $(Y_i - d_i)/(Y_i - 1)$, que es igual a uno si no hay dos individuos que tengan un tiempo de evento común, es una corrección en caso de empates. Los términos $(Y_{ij}/Y_i)(1 - Y_{ij}/Y_i)d_i$ y $-(Y_{ij}/Y_i)(Y_{ig}/Y_i)d_i$ surgen de la varianza y covarianza de una variable aleatoria multinomial con parámetros d_i, $p_j = Y_{ij}/Y_i, j = 1, ..., K$.

Klein y Moeschberger [6] mencionan también que el vector de componentes $(Z_1(\tau), ..., Z_K(\tau))$ son linealmente dependientes porque $\sum_{j_1}^{K} Z_j(\tau)$ es cero. El estadístico de la prueba se construye seleccionando cualquier $K - 1$ de las Z_j. La matriz de varianza-covarianza estimada de estos estadísticos viene dada por la matriz $(K - 1) \times (K - 1)$ Σ , formada por los $\hat{\sigma}_{jg}$ apropiados. El estadístico de la prueba viene dada por la forma cuadrática

$$\chi^2 = (Z_1(\tau), ..., Z_K(\tau))\Sigma^{-1}(Z_1(\tau), ..., Z_K(\tau)^t \tag{2.21}$$

Cuando la hipótesis nula es verdadera, este estadístico tiene una distribución chi-cuadrado, para muestras grandes con $K - 1$ grados de libertad. Una prueba de nivel α de H_0 se rechaza cuando χ^2 es mayor que el α-ésimo punto porcentual superior de una variable aleatoria de chi-cuadrado con $K - 1$ grados de libertad.

Klein y Moeschberger [6] también estudian el caso particular $K = 2$ (que es el que en particular nos interesa para este trabajo). Cuando $K = 2$, el estadístico de prueba se puede escribir como

$$Z = \frac{\sum_{i=1}^{D} W(t_i)[d_{i1} - Y_{i1}\frac{di}{Yi}]}{\sqrt{\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i}\right) \left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i}} \tag{2.22}$$

que tiene una distribución normal estándar para muestras grandes cuando H_0 es verdadera. Usando este estadístico, una prueba de nivel α de la hipótesis alternativa $H_A : \ h_1(t) > h_2(t)$, para algún $t \leq \tau$, se rechaza cuando $Z \geq Z_\alpha$, el α-ésimo punto porcentual superior de una distribución normal estándar. La prueba de $H_A : \ h_1(t) \neq h_2(t)$, para algún t, se rechaza cuando $|Z| > Z_{\alpha/2}$.

Klein y Moeschberger [6] también comentan que en la literatura se ha propuesto una variedad de funciones de ponderación. Una función de peso común, que conduce a una prueba disponible en la mayoría de los paquetes estadísticos, es $W(t) = 1$ para todo t. Esta elec-

ción de función de ponderación conduce a la llamada prueba log-rank y tiene un poder óptimo para detectar alternativas donde las tasas de riesgo en las K poblaciones son proporcionales entre sí. Una segunda elección de pesos es $W(t_i) = Y_i$. Esta función de ponderación produce la generalización de Gehan [9] de la prueba de Mann-Whitney-Wilcoxon de dos muestras y la generalización de Breslow [10] de la prueba Kruskal-Wallis. Tarone y Ware [11] sugieren una clase de pruebas donde la función de peso es $W(t_i) = f(Y_i)$ y f es una función fija. Sugieren la elección de $f(y) = y^{1/2}$. Esta clase de ponderaciones da más peso a las diferencias entre el número de muertes observadas y esperadas en la muestra j en los puntos temporales donde hay más datos.

Peto y Peto [12] propusieron una versión alternativa de datos censurados de la prueba de Mann-Whitney-Wilcoxon. Se define una estimación de la función de supervivencia común por

$$\widetilde{S} = \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i + 1}\right) \tag{2.23}$$

que está cerca del estimador Producto-Límite. Sugieren usar $W(t_i) = \widetilde{S}(t_i)$. Andersen y col. [13] sugieren que este peso debería modificarse ligeramente como $W(ti) = \widetilde{S}(t_i)Y_i/(Y_i+1)$. Cualquiera de las ponderaciones depende de la experiencia de supervivencia combinada en la muestra combinada, mientras que la ponderación $W(t_i) = Y_i$ depende en gran medida de los tiempos de los eventos y las distribuciones de censura. Debido a este hecho, las ponderaciones de Gehan-Breslow pueden tener resultados engañosos cuando los patrones de censura son diferentes en las muestras individuales.

Fleming y Harrington [14] proponen una clase de pruebas muy generales que incluye, como casos especiales, la prueba de log-rank y una versión de la prueba de Mann-Whitney-Wilcoxon, muy cercana a la sugerida por Peto y Peto [12]. Sea $\widetilde{S}(t)$ el estimador Producto-Límite basado en la muestra combinada. Su función de peso está dada por

$$W_{p,q}(t_i) = \widetilde{S}(t_{i-1})^p [1 - \widetilde{S}(t_{i-1})]^q,\ p \geq 0,\ q \geq 0 \tag{2.24}$$

En este caso, la función de supervivencia en el momento de la muerte anterior se utiliza como ponderación para garantizar que estas ponderaciones se conozcan justo antes del momento en el que se realizará la comparación. Note que $S(t_0) = 1$ y definimos $0^0 = 1$ para estos pesos. Cuando $p = q = 0$ para esta clase, tenemos la prueba de rango logarítmico. Cuando $p = 1$, $q = 0$, tenemos una versión de la prueba de Mann-Whitney-Wilcoxon. Cuando $q = 0$ y $p > 0$, estas ponderaciones dan mayor peso a las salidas tempranas entre las

tasas de riesgo en las K poblaciones, mientras que, cuando $p = 0$ y $q > 0$, estas pruebas dan mayor peso a las salidas que ocurren tarde en el tiempo. Mediante una elección apropiada de p y q, se pueden construir pruebas que tienen el mayor poder contra las hipótesis alternativas que tienen las K tasas de riesgo que difieren en cualquier región deseada.

2.2 EL MODELO DE COX

2.2.1 FORMULACIÓN DEL MODELO

El modelo de Cox [15] enuncia la función de riesgo h en función del tiempo t y de un conjunto de variables explicativas[1], $X = (x_1, \ldots, x_p)$, a través de la cual el sujeto en estudio es definido de la siguiente manera:

$$h\left(t, X\right) = h\left(t, x_1, \ldots, x_p\right) = h_0\left(t\right)\ exp\left(\sum_{j=1}^{p} \beta_j x_j\right) \tag{2.25}$$

Los supuestos del modelo de Cox son:

1. Que la razón de riesgo es constante en el tiempo (llamada la **hipótesis de riesgos proporcionales**, es el eje fundamental del modelo de Cox, por lo que se dedicará un apartado para explicarla en detalle).
2. Que las covariables continuas tienen una forma funcional adecuada.
3. Que los sujetos no tienen influencia en la estimación de cada coeficiente.
4. Que los sujetos no tienen influencia en la estimación del modelo (No existen casos atípicos que hagan que el modelo sufra desvíos).
5. Que no existe heterogeneidad no observada.
6. *Que los tiempos de supervivencia tienen distribuciones continuas.*
7. *Que los tiempos están tomados de forma exacta y que no existe posibilidad de empates.*

[1] También llamadas covariables, predictores, factores de riesgo, variables de confusión.

Aunque el modelo original implica el cumplimiento de los supuestos 6 y 7, la verdad es que se han desarrollado alternativas que no requieren el cumplimiento de los mismos, por lo que en realidad no son estrictamente necesarios, como veremos después.

Luego entonces, para cada sujeto i con $i \in \{1, \ldots, n\}$ conoceremos su tiempo de muerte/fallo[2] t_i , su estado de fallo o censura d_i , variable codificada con 1 si el dato no está censurado y con 0 si el dato sí lo está[3], y las covariables fijas $X_i = (x_{i1}, \ldots, x_{ip})$. Si incluimos el subíndice i para denotar a un sujeto determinado, el modelo 2.25 se podría reescribir como:

$$h\left(t_i, X_i\right) = h\left(t_i, x_{i1}, \ldots, x_{ip}\right) = h_0\left(t_i\right)\ exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right).$$

La función $h_0\left(t\right)$ es llamada "función de riesgo basal" y corresponde al riesgo de un individuo que tiene como valor 0 (cero) en todos los factores de riesgo. Éste sería el "individuo de referencia" de cara a la posterior interpretación del análisis:

$$h\left(t, x_1 = 0, \ldots, x_p = 0\right) = h_0\left(t\right)\ exp\left(\sum_{j=1}^{p} \beta_j 0\right) = h_0\left(t\right).$$

Podríamos decir que la función de riesgo basal es aquella función "básica" del modelo si éste no incorporara predictores.

La función de riesgo basal, $h_0\left(t\right)$, es la única parte de la expresión del modelo de Cox que depende del tiempo t. La otra parte, $\exp\left(\sum_{j=1}^{p} \beta_j x_j\right)$, sólo depende del vector de covariables $X = (x_1, \ldots, x_p)$ de los sujetos. Para este apartado supondremos que los predictores son "independientes del tiempo".

Por definición, una variable independiente del tiempo es aquella cuyo valor no varía a lo largo del tiempo, por ejemplo, el sexo, la raza o el antibiótico que se usa (Cuando hablamos de estudios médicos) que son variables fijas que sólo toman un valor, el inicial. También se pueden considerar algunas variables, como el hecho de ser o no fumador (condición de

[2]Tiempo que transcurre hasta que ocurre el evento de interés.

[3]El software R usa esta codificación, sin embargo, es importante tomar en cuenta que R toma, como ya se dijo, 0 para censura a la izquierda (El evento no ocurrió durante el periodo experimental) y 2 para censura a la derecha (El evento ocurrió antes del inicio del experimento).

fumador), como independientes del tiempo, ya que, aunque la condición de fumador puede variar en el tiempo, se suele suponer que para el estudio no varía, se parte de un estado inicial y se considera que no cambiará hasta el final, y por lo tanto que sólo toma un valor por individuo. Otro ejemplo de este tipo podría ser la variable "estado inicial de la enfermedad".

Cabe recalcar que existen variables que, aunque en efecto varían en función del tiempo, se suelen tratar como independientes del tiempo. Un ejemplo de estas variables son la edad y el peso de los sujetos que, si bien varían con el tiempo, puede ser apropiado tratarlas como independientes del tiempo en análisis determinados; esto es posible siempre que los valores de estos predictores varíen mínimamente a lo largo del tiempo o también si el impacto de dichas variables en el riesgo de supervivencia depende en esencia de un único valor de medición.

Adicionalmente, es posible incluir predictores dependientes del tiempo a los que denominaremos $X(t) = (x_1(t), \ldots, x_p(t))$. Por ejemplo, el estado corriente de la enfermedad o medidas de tensión arterial sucesivas. En tal caso es posible utilizar la modelo de Cox pero es común que no se satisfaga el requisito de "riesgos proporcionales" que más abajo se definirá. En esta situación en que se tienen en cuenta predictores que dependen del tiempo la regresión se denomina "modelo de Cox ampliado". Este modelo no será explicado con el objetivo de enfocarnos únicamente en el modelo de Cox que satisface la condición de riesgos proporcionales.

Se considera al modelo de Cox un modelo "semiparamétrico"[16] debido a que consta de una parte paramétrica y una parte nó paramétrica:

1. La parte "paramétrica" está formada por la expresión $\exp\left(\sum_{j=1}^{p} \beta_j x_j\right)$, es decir, la exponencial del predictor lineal $\eta = \sum_{j=1}^{p} \beta_j x_j$. En esta parte del modelo se estiman los parámetros (o coeficientes del predictor lineal) $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ de la regresión mediante la maximización de la denominada "función de verosimilitud parcial" que veremos detalladamente en un apartado posterior.

2. La parte "no paramétrica" está formada por la función de riesgo basal $h_0(t)$ que es una función arbitraria, no especificada y estimada en una segunda etapa condicionada a la estimación de los parámetros $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ de la regresión. El modelo de Cox se considera "semiparamétrico" debido a esta componente no paramétrica de la fórmula.

Ya estimadas la parte paramétrica y, posteriormente, la no paramétrica tenemos el modelo

semiparamétrico en su totalidad:

$$\hat{h}\left(t, X\right) = \hat{h}_0\left(t\right)\ exp\left(\sum_{j=1}^{p} \hat{\beta}_j x_j\right)$$

Como ya lo dijimos anteriormente, La función de riesgo basal $h_0\left(t\right)$ no está especificada en el modelo de Cox y por lo tanto la distribución del error tampoco. Esta particularidad de que el modelo de Cox sea un modelo "semiparamétrico" hace que sea bien recibido en análisis de supervivencia. Una función de riesgo basal no especificada permite estimar los coeficientes de la regresión, calcular las razones de riesgo y ajustar las curvas de supervivencia a una gran variedad de situaciones.

Como los ajustes resultantes tienden a aproximarse al modelo paramétrico adecuado, podemos decir que el modelo de Cox es "robusto" , al menos en este sentido. A manera de ejemplo, supongamos que el modelo paramétrico correcto para un estudio es el Weibull, entonces las curvas de supervivencia que se conseguirán con el modelo de Cox serán similares a las obtenidas con el modelo paramétrico adecuado.

Siempre será preferible usar el modelo paramétrico que se ajuste adecuadamente al estudio antes que el modelo de Cox, sin embargo, no siempre podemos saber que modelo es apropiado y pocas veces podemos tener modelos paramétricos que se ajusten adecuadamente (Esto se puede medir con una prueba de bondad de ajuste). Por lo anterior, el modelo de Cox resulta ser una buena alternativa sobre todo si estamos entrando en terreno desconocido, como es nuestro caso, y nos evitará caer en errores como puede ser, por ejemplo, un modelo paramétrico sobre entrenado que se ajusta bien con los datos de entrenamiento pero que no da buenos resultados con los datos de prueba.

La metodología para el tratamiento y codificación de predictores en el modelo de Cox de riesgos proporcionales es análoga a la usada en los modelos lineales o lineales generalizados, con excepción de término "intercept" o medio (la constante de la expresión lineal) que queda absorbido por la función de riesgo basal. Se pueden tener en cuenta los efectos principales y también interacciones.

2.2.2 HIPÓTESIS DE RIESGOS PROPORCIONALES

Para el modelo de Cox buscamos en primer lugar la relación entre los riesgos de muerte de dos individuos expuestos a factores de riesgo diferentes; por esta razón, el modelo parte de una hipótesis fundamental, la de que los riesgos son proporcionales.

Previo a explicar a qué se refiere esta hipótesis fundamental, debemos definir lo que es la "razón de riesgos" (traducción del inglés "Hazard Ratio").

La razón de riesgos (HR) entre dos sujetos con vectores de covariables $X = (x_1, \ldots, x_p)$ y $X^* = \left(x_1^*, \ldots, x_p^*\right)$, respectivamente, se define como:

$$HR = \frac{h\left(t, X^*\right)}{h\left(t, X\right)} \tag{2.26}$$

Es común o típico evaluar en el numerador el individuo que tiene mayor riesgo definido por X^* y en el denominador el sujeto de menor riesgo definido por X , de la misma manera que se hace con los $odds\ ratio$. Por obvias razones se espera que $h\left(t, X^*\right) > h\left(t, X\right)$ y por lo tanto que el HR sea mayor que 1, lo que nos permitiría cuantificar cuantas veces es mayor el riesgo del individuo definido por X^* con respecto al individuo X . La razón de riesgos es mucho más fácil de interpretar y explicar cuando su valor es mayor que 1 que cuando es menor que la unidad.

Teniendo claro lo anterior podemos seguir con el análisis sustituyendo la expresión 2.25 en 2.26 , lo que nos da el siguiente resultado:

$$HR = \frac{h_0\left(t\right)\ exp\left(\sum_{j=1}^{p} \beta_j x_j^*\right)}{h_0\left(t\right)\ exp\left(\sum_{j=1}^{p} \beta_j x_j\right)} = \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_j^*\right)}{\exp\left(\sum_{j=1}^{p} \beta_j x_j\right)} = \exp\left(\sum_{j=1}^{p} \beta_j \left(x_j^* - x_j\right)\right),$$

con lo que tenemos

$$HR = \exp\left(\sum_{j=1}^{p} \beta_j \left(x_j^* - x_j\right)\right) \tag{2.27}$$

Podemos ver que la razón de riesgos no depende de la función de riesgo basal, sino únicamente de los factores de riesgo y de los coeficientes $\beta_1, \ldots, \beta_p$, es decir, no depende del tiempo. En conclusión, el modelo de Cox supone la hipótesis de que los riesgos son

proporcionales ya que las covariables no son dependientes del tiempo.

La hipótesis de riesgos proporcionales significa, explícitamente, que la razón de riesgo 2.26 es constante en el tiempo, lo que se traduce en: $h\left(t, X^{*}\right) = constante \times h\left(t, X\right)$. Denominaremos a la constante θ , luego, aplicando la expresión 2.27 y una vez estimados los coeficientes de la regresión por máxima verosimilitud parcial, tenemos que la razón de proporcionalidad es constante en el tiempo e igual a:

$$\hat{\theta} = exp\left(\sum_{j=1}^{p} \hat{\beta}_j \left(x_j^{*} - x_j\right)\right) \tag{2.28}$$

Sean dos individuos i y j con factores de riesgo son $X^i = \left(x_1^i, \ldots, x_p^i\right)$ y $X^j = \left(x_1^j, \ldots, x_p^j\right)$, respectivamente, tal que $x_u^i = x_u^j = x_u$ para todo $u \in \{1, \ldots, p\} / \{k\}$, $x_k^i = 1$ y $x_k^j = 0$; tenemos que para cualquier tiempo t:

$$HR = \frac{h\left(t, x_1^i, \ldots, x_k^i, \ldots, x_p^i\right)}{h\left(t, x_1^j, \ldots, x_k^j, \ldots, x_p^j\right)} = \frac{h\left(t, x_1, \ldots, 1, \ldots, x_p\right)}{h\left(t, x_1, \ldots, 0, \ldots, x_p\right)}$$

$$HR = \frac{h_0\left(t\right)\ exp\left(\beta_1 x_1 + \ldots + \beta_k 1 + \ldots + \beta_p x_p\right)}{h_0\left(t\right)\ exp\left(\beta_1 x_1 + \ldots + \beta_k 0 + \ldots + \beta_p x_p\right)} = \frac{\exp\left(\beta_1 x_1 + \ldots + \beta_k + \ldots + \beta_p x_p\right)}{\exp\left(\beta_1 x_1 + \ldots + 0 + \ldots + \beta_p x_p\right)}$$

$$HR = exp\left(\beta_k\right)$$

Para el caso de factores de riesgo continuos tenemos que $\exp\left(\beta_l\right)$ representa la razón de riesgo al incrementar en una unidad el factor de riesgo x_l . De la misma manera, si incrementamos la covariable x_l en c unidades, entonces la razón de riesgo para ese caso será $\exp\left(c\beta_l\right)$.

Por su puesto, cuando utilicemos el modelo de Cox es necesario verificar que se cumple la hipótesis de riesgos proporcionales. Es común que se compruebe la hipótesis al verificar que el impacto de cada variable es constante en el tiempo. Existen varios métodos para verificar el cumplimiento de esta hipótesis, mencionaremos dos:

- ❒ Método gráfico. - Cuando una variable toma únicamente valores 0 y 1 (o toma solamente dos valores, no necesariamente 0 y 1), podemos representar las gráficas de

supervivencia de cada grupo y verificar si son paralelas.

- ❒ Residuos de Schoenfeld. - El estudio de los residuos de Schoenfeld es un método estadístico mucho más riguroso, por lo que se le dedicará algunos párrafos para su estudio.

2.2.3 FUNCIÓN DE VEROSIMILITUD PARCIAL Y ESTIMACIÓN DE LOS COEFICIENTES

Los parámetros $\beta = (\beta_1, \ldots, \beta_p)$ se estiman maximizando el logaritmo de la llamada "función de verosimilitud parcial" mediante métodos numéricos, de esta manera obtenemos la estimación $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$. Con la estimación de estos parámetros ya tendremos la componente paramétrica totalmente definida en el modelo de Cox y en consecuencia podremos hacer inferencia sobre dicho vector de parámetros y calcular las razones de riego de interés para el estudio.

Debido a la existencia de datos incompletos, los parámetros del modelo de Cox no pueden ser estimados por el método ordinario de máxima verosimilitud al ser desconocida la forma específica de la función arbitraria de riesgo. Cox propuso un método de estimación denominado verosimilitud parcial [17], siendo las verosimilitudes condicionales y marginales casos particulares del anterior. El método de verosimilitud parcial se diferencia del método de verosimilitud ordinario en el sentido de que mientras el método ordinario se basa en el producto de las verosimilitudes para todos los individuos de la muestra, el método parcial se basa en el producto de las verosimilitudes de todos los cambios ocurridos, es decir, no toma en cuenta las verosimilitudes de los datos censurados, sin embargo, en el cálculo de las probabilidades de los tiempos de supervivencia sí se tiene en cuenta a todos los sujetos (censurados o no a posteriori) objeto de riesgo al inicio de los diferentes tiempos de ocurrencia del evento de interés.

Definimos $L \equiv L(\beta_1, \ldots, \beta_p)$ como la función de verosimilitud parcial. Sean n individuos de los cuales k son datos no censurados, es decir, tenemos k tiempos de ocurrencia del evento de interés y $n-k$ datos censurados, notaremos los tiempos de cada individuo, censurado o no, por $t_1, \ldots, t_n$, los tiempos de ocurrencia del evento de interés ordenados por $t_{(1)}, \ldots, t_{(k)}$, solo para individuos no censurados, y al conjunto de los sujetos a riesgo al tiempo $t_{(i)}$ por $R\left(t_{(i)}\right)$ para todo $i \in \{1, \ldots, k\}$, es decir, $R\left(t_{(i)}\right) = \left\{v \in \{1, \ldots, n\} \,|\, t_v \geq t_{(i)}\right\}$. También

definiremos $L_i \equiv L_{t_{(i)}}(\beta_1, \ldots, \beta_p)$ para $i \in \{1, \ldots, k\}$ a las porciones de la verosimilitud total anterior que se deben a los aportes de los diferentes tiempos de ocurrencia del evento de interés.

Se construye la función de verosimilitud total como el producto de las aportaciones de cada uno de los k tiempos de ocurrencia del evento de interés:

$$L = \prod_{i=1}^{k} L_i$$

Una vez construida la función de verosimilitud total, tomamos el logaritmo de la misma y derivamos respecto de cada parámetro:

$$\frac{\partial \log L}{\partial \beta_j} \tag{2.29}$$

$$\frac{\partial^2 \log L}{\partial \beta_i \partial \beta_j} \tag{2.30}$$

Si para todo $j \in \{1, \ldots, p\}$ igualamos la expresión 2.29 a 0, obtenemos las ecuaciones que nos permiten calcular los estimadores $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ mediante algún método numérico.

Además, a partir de 2.30 podemos comprobar que realmente es un máximo y obtener la matriz de información (observada), $I(\beta)$, donde cada elemento de la matriz de $p \times p$ es igual a:

$$I_{ij}(\beta) = -\frac{\partial^2 \log L}{\partial \beta_i \partial \beta_j}$$

de manera que la matriz de varianzas y covarianzas estimada es $\hat{\Sigma} = I^{-1}\left(\hat{\beta}\right)$ Este estimador es asintóticamente no sesgado, eficiente y normal.

Es necesario mencionar que, aunque $\hat{\beta}$ estima consistentemente el vector de parámetros β , no es completamente eficiente, es decir, no alcanza la cota de Cramer-Rao; otra de las características de este estimador es que tiene una distribución aproximadamente normal de media $\left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ y matriz de varianzas y covarianzas $\hat{\Sigma}$ **.**

Podemos calcular las $L_i \equiv L_{t_{(i)}}(\beta_1, \ldots, \beta_p)$ para todo i tal que $i \in \{1, \ldots, k\}$ mediante:

$$L_{t_{(i)}}(\beta_1,\ldots,\beta_p) = \frac{h\left(t_{(i)}, X_{(i)}\right)}{\sum_{l\in R\left(t_{(i)}\right)} h\left(t_{(i)}, X_l\right)} = \frac{h_0\left(t_{(i)}\right)\exp\left(\sum_{j=1}^{p}\beta_j X_{ij}\right)}{\sum_{l\in R\left(t_{(i)}\right)} h_0\left(t_{(i)}\right)\exp\left(\sum_{j=1}^{p}\beta_j X_{lj}\right)}$$

$$L_{t_{(i)}}(\beta_1,\ldots,\beta_p) = \frac{\exp\left(\sum_{j=1}^{p}\beta_j X_{ij}\right)}{\sum_{l\in R\left(t_{(i)}\right)} \exp\left(\sum_{j=1}^{p}\beta_j X_{lj}\right)} \tag{2.31}$$

Donde $X_{(i)}$ es el vector de covariables del sujeto con tiempo de fallo $t_{(i)}$ y X_l para $l \in R\left(t_{(i)}\right)$ el vector de factores de riesgo de cada uno de los sujetos de $R\left(t_{(i)}\right)$ según fue definido anteriormente.

Tal como hemos definido la función de verosimilitud, ésta no depende de los tiempos sino únicamente de su orden y de si el dato estaba o no censurado. En consecuencia, podemos tener las mismas estimaciones de $\hat{\beta}$ para distintos tiempos, siempre que tengan el mismo orden y censura de datos.

2.2.4 FUNCIÓN DE VEROSIMILITUD EN CASO DE EMPATES

Para poder tener en cuenta los empates en los tiempos de supervivencia podemos usar tres métodos, todos estos aplicables mediante la función ‘coxphíncluida en el paquete survival[18] del software R[19]:

- ❒ Método de Breslow (ties “breslow”)
- ❒ Método de Efron (ties “efron”)
- ❒ Método “Exact partial likelihood” (ties “exact”)

Por defecto, se usa el método de Efron para esta función, ya que según el manual es la opción computacionalmente más eficiente para tiempos continuos. El método “exact” es equivalente a un modelo de regresión logística condicional y podría ser apropiado cuando los tiempos sean un conjunto discreto y relativamente pequeño. En este último caso, si los datos tienen tamaño “grande” el tiempo de cómputo puede ser excesivo. Por otro lado, cuando tenemos un número elevado de empates, la aproximación discreta es, como veremos a continuación, computacionalmente complicada.

Añadiremos, a partir de aquí, una nomenclatura para poder tener las fórmulas de manera

explícita. Empecemos por escribir la expresión 2.31 de forma matricial asumiendo vectores fila para los parámetros y las covariables:

$$L_{t_{(i)}}(\beta) = \frac{\exp\left(\beta X'_{(i)}\right)}{\sum_{l \in R\left(t_{(i)}\right)} \exp\left(\beta X'_l\right)} \tag{2.32}$$

Con esta misma notación, la función de verosimilitud parcial total será:

$$L(\beta) = \prod_{i=1}^{k} L_{t_{(i)}}(\beta) = \prod_{i=1}^{k} \frac{\exp\left(\beta X'_{(i)}\right)}{\sum_{l \in R\left(t_{(i)}\right)} \exp\left(\beta X'_l\right)} \tag{2.33}$$

Notaremos además $m_{(i)}$ a la multiplicidad de eventos de $t_{(i)}$ para $i \in \{1, \ldots, k\}$, es decir, el número de sujetos que tienen el mismo tiempo de fallo $t_{(i)}$. También notaremos $r_i = card\left[R\left(t_{(i)}\right)\right]$, es decir, el número de sujetos en riesgo al tiempo $t_{(i)}$.

Por ejemplo, sea el conjunto de tiempos de supervivencia experimentales $\{15, 16^*, 20, 20, 20, 21, 24, 24\}$ donde los valores señalados con asterisco ($*$) son datos censurados, entonces tenemos que: $n = 8, k = 4, t_{(1)} = 15, t_{(2)} = 20, t_{(3)} = 21, t_{(4)} = 24, m_{(1)} = 1, m_{(2)} = 3, m_{(3)} = 1, m_{(4)} = 2, r_1 = 8, r_2 = 6, r_3 = 3, r_4 = 2$.

Además, para todo $R\left(t_{(i)}\right)$ es posible seleccionar aleatoriamente $m_{(i)}$ sujetos; notaremos cada una de estas posibles selecciones por $u_{(j)}$ y U_i el conjunto de todas estas, de manera que tendremos $C^{r_i}_{m_{(i)}} = \frac{r_i!}{\left[m_{(i)}!\left(r_i - m_{(i)}\right)!\right]}$ posibles $u_{(j)}$'s. En nuestro ejemplo, para $R\left(t_{(2)}\right)$ podemos seleccionar aleatoriamente 3 sujetos de entre 6 ($m_{(2)} = 3, r_2 = 6$), de manera que tendremos $C^6_3 = 20$ posibles combinaciones y $U_2 = \left\{u_{(1)}, \ldots, u_{(20)}\right\}$.

Tomemos ahora los sujetos empatados. Sea $X_k = (x_{k1}, \ldots, x_{kp})$ el vector de covariables del k-ésimo sujeto, diremos que $Z_{u_{(j)}} = \sum_{k \in u_{(j)}} X_k = \left(z_{u_{(j)}1}, \ldots, z_{u_{(j)}p}\right)$, donde $z_{u_{(j)}l}$ es la suma de la l-ésima covariable de los $m_{(i)}$ sujetos que están en $u_{(j)}$. Luego, sea $u^*_{(j)}$ el conjunto de los $m_{(i)}$ sujetos a quienes ocurre el evento de interés al tiempo $t_{(i)}$ (sujetos empatados) y $Z^*_{u_{(j)}} = \sum_{k \in u^*_{(j)}} X_k = \left(z_{u^*_{(j)}1}, \ldots, z_{u^*_{(j)}p}\right)$, donde $z_{u^*_{(j)}l}$ es la suma de la l-ésima covariable de los $m_{(i)}$ sujetos que están en $u^*_{(j)}$. Tomando el mismo ejemplo de antes, $z_{u^*_{(2)}1}$ es la suma de la primera componente de las covariables de los 3 sujetos que tienen el tiempo de fallo 20, a saber, el tercero, el cuarto y el quinto sujetos.

2.2.4.1 Tiempo Continuo

Para tiempo continuo, cuando $m_{(i)}$ es pequeño en comparación de r_i (por lo general), tenemos dos opciones la aproximación de Breslow o la de Efron, expresadas a continuación:

- ❒ Breslow [20]:

$$L_E(\beta) = \prod_{i=1}^{k} \frac{\exp\left(\beta Z'_{u^*_{(i)}}\right)}{\left[\sum_{l \in R(t_{(i)})} \exp\left(\beta X'_l\right)\right]^{m_{(i)}}} \tag{2.34}$$

- ❒ Efron [21]:

$$L_E(\beta) = \prod_{i=1}^{k} \frac{\exp\left(\beta Z'_{u^*_{(i)}}\right)}{\prod_{j=1}^{m_{(i)}} \left[\sum_{l \in R(t_{(i)})} \exp\left(\beta X'_l\right) - \left[\frac{(j-1)}{m_{(i)}}\right] \sum_{l \in u^*_{(i)}} \exp\left(\beta X'_l\right)\right]} \tag{2.35}$$

2.2.4.2 Tiempo Discreto

Cuando el tiempo es una variable discreta los empates son "verdaderos", es decir, los fallos verdaderamente ocurren al mismo tiempo. Para casos como este, Cox[15] propuso el siguiente modelo:

$$\frac{h(t, X)\, dt}{1 - h(t, X)\, dt} = \frac{h_0(t)\, dt}{1 - h_0(t)\, dt} \exp\left(\beta X'\right)$$

En una escala continua para el tiempo, este modelo se reduce a 2.25. En la función de verosimilitud parcial, reemplazamos en 2.32 el siguiente término con empates:

$$L_{t_{(i)}}(\beta) = \frac{\exp\left(\beta Z'_{u^*_{(i)}}\right)}{\sum_{u_{(j)} \in U_i} \exp\left(\beta Z'_{u_{(j)}}\right)}$$

Con lo que finalmente tenemos que la función de verosimilitud parcial total para el caso de empates en tiempo discreto será:

$$L_D(\beta) = \prod_{i=1}^{k} \frac{\exp\left(\beta Z'_{u^*_{(i)}}\right)}{\sum_{u_{(j)} \in U_i} \exp\left(\beta Z'_{u_{(j)}}\right)} \tag{2.36}$$

Donde el i-ésimo termino nos dice la probabilidad condicional de observar los $m_{(i)}$ fallos al tiempo $t_{(i)}$ y en el conjunto de riesgo $R\left(t_{(i)}\right)$. El número de términos en el denominador vendría a ser $C^{r_i}_{m_{(i)}} = \frac{r_i!}{\left[m_{(i)}!\left(r_i - m_{(i)}\right)!\right]}$.

Para solucionar la ecuación 2.36 se utiliza un algoritmo recurrente, lo que hace que el tiempo computacional sea considerablemente menor. Se puede considerar 2.36 como una aproximación de la función de verosimilitud parcial total para tiempos de supervivencia continuos con empates suponiendo que los empates son verdaderos como si se hubieran observado en una escala discreta de tiempo.

En la práctica, las ecuaciones 2.34, 2.35 y 2.36 son aproximaciones cercanas de la función de verosimilitud total exacta para tiempos de supervivencia en tiempo continuo con empates. Los tres casos se reducen a 2.33 cuando no hay empates.

2.2.5 CONTRATES DE HIPÓTESIS

Mediante la función de verosimilitud obtenemos una estimación $\hat{\beta}$ de los parámetros, la cual tiene una distribución aproximadamente normal de media $\left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ y matriz de varianzas y covarianzas $\hat{\Sigma}$; esto da paso al uso de pruebas de hipótesis análogas a las que se usa en modelos lineales generalizados.

Si lo que queremos es demostrar la hipótesis nula $H_0 : \ \beta_j = 0$ contra la hipótesis alternativa $H_1 : \ \beta_j \neq 0$, podemos resolverlo utilizando el estadístico de Wald:

$$z = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}\left(\hat{\beta}_j\right)}}$$

El intervalo de confianza aproximado de nivel $(1 - \alpha)$ para el j-ésimo coeficiente se calcula mediante:

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}\left(\hat{\beta}_j\right)}$$

Si en cambio queremos demostrar la hipótesis nula $H_0 : \ \beta = \beta_0$ contra la hipótesis alternativa $H_1 : \ \beta \neq \beta_0$, podemos utilizar tres pruebas:

1. **Contraste de Wald**. Este contraste se basa en que $\hat{\beta}$ sigue asintóticamente una dis-

tribución aproximadamente normal de media $\left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ y matriz de varianzas y covarianzas estimada es $\hat{\Sigma} = I^{-1}\left(\hat{\beta}\right)$. Se define como:

$$\chi_W = \left(\hat{\beta} - \hat{\beta}_0\right)^T I\left(\hat{\beta}\right)\left(\hat{\beta} - \hat{\beta}_0\right)$$

Bajo la hipótesis nula, este estadístico sigue una distribución χ^2 con p grados de libertad.

2. **Contraste de la razón de verosimilitudes**. En este contraste se utiliza el valor de la función de verosimilitud parcial evaluada en $\hat{\beta}$ **,** $L\left(\hat{\beta}\right)$ y se la compara con la verosimilitud parcial evaluada bajo la hipótesis nula β_0 , $L(\beta_0)$. Se define como:

$$\chi_{LR} = 2\left(\log L\left(\hat{\beta}\right) - \log L(\beta_0)\right)$$

Bajo la hipótesis nula, este estadístico sigue una distribución χ^2 con p grados de libertad.

3. **Contraste del "score" (Log Rank).** En este contraste se utiliza el gradiente del logaritmo de la verosimilitud parcial evaluada en la hipótesis nula y asume que bajo la hipótesis nula el vector scores:

$$\chi_S = \left(\frac{\partial L(\beta_0)}{\partial \beta}\right)^T \left(-\frac{\partial L^2(\beta_0)}{\partial \beta \partial \beta^T}\right)^{-1} \frac{\partial L(\beta_0)}{\partial \beta}$$

Bajo la hipótesis nula, este estadístico sigue una distribución χ^2 con p grados de libertad.

2.2.6 AJUSTE DE LAS CURVAS DE RIESGO O SUPERVIVENCIA EN EL MODELO DE COX

El modelo de Cox se puede expresar en términos de funciones de supervivencia, ya que se cumple la relación $S(t) = exp\left(\int_0^t h(s)\, ds\right)$, donde h se define como en 2.25 . De esta manera es fácil comprobar que el modelo de Cox en función de las curvas de supervivencia es:

$$S(t, X) = S_0(t)^{\exp\left(\sum_{j=1}^p \beta X\right)} \tag{2.37}$$

donde $S_0(t)$ se denomina la función de supervivencia basal siguiendo la idea de la función de riesgo basal, $h_0(t)$, que corresponde a la supervivencia de un sujeto base (cuyas covariables son todas iguales a cero) al tiempo t.

Como ya lo dijimos en párrafos anteriores, en el modelo de Cox se ajustan las curvas de supervivencia tomando en cuenta los coeficientes ya estimados (la parte paramétrica del modelo), por lo que se tiene en cuenta los valores de las variables explicativas (covariables) para el ajuste de las curvas. A estas curvas las denominamos "curvas de supervivencia ajustadas" y de la misma manera que las curvas Kaplan-Meier se trazan como funciones escalonadas para cada uno de los tiempos de supervivencia. Las curvas estimadas se definen como:

$$\hat{S}(t, X) = \hat{S}_0(t)^{\exp\left(\sum_{j=1}^{p} \hat{\beta} X\right)} \tag{2.38}$$

En esta parte del proceso, se estima la parte no paramétrica que se relaciona directamente a la función de supervivencia basal. De esta manera, la estimación de la función de supervivencia dependerá de los tiempos de supervivencia, los valores de las variables explicativas del modelo y la estimación de la parte paramétrica del modelo.

A continuación, presentaremos tres métodos de estimación de esta función no paramétrica y no especificada en caso de empates[4]:

- ❒ Método Nelson-Aalen-Breslow
- ❒ Método Efron[5]
- ❒ Método Kalbfleish-Prentice

Estos métodos de estimación de la parte no paramétrica del modelo Cox se basan en metodologías de maximización de una función de verosimilitud definida con las aportaciones de las curvas de supervivencia, de manera que, bajo condiciones de regularidad, las curvas siguen una distribución normal, de manera que es posible calcular fácilmente la esperanza y la varianza. De todo lo anterior se sigue que podemos calcular intervalos de confianza.

[4]Estos tres métodos son los mismos que contempla el paquete survival del software R.

[5]Método por defecto en el paquete survival del software R.

2.2.6.1 Método Nelson-Aalen-Breslow

[22]

$$\hat{H}_0(t) = \sum_{t_{(i)} \leq t} \left[\frac{m_{(i)}}{\sum_{l \in R(t_{(i)})} exp\left(\hat{\beta} X_l'\right)} \right] \tag{2.39}$$

donde $\hat{S}_0 = exp\left[-\hat{H}_0(t)\right]$ y las curvas estimadas serán 2.38.

2.2.6.2 Método Efron

[23]

$$\hat{H}_0(t) = \sum_{t_{(i)} \leq t} \left[\sum_{j=1}^{m_{(i)}} \frac{1}{\sum_{l \in R(t_{(i)})} exp\left(\hat{\beta} X_l'\right) - \left[(j-1)/m_{(i)}\right] \sum_{l \in u^*_{(i)}} exp\left(\hat{\beta} X_l'\right)} \right] \tag{2.40}$$

donde $\hat{S}_0 = exp\left[-\hat{H}_0(t)\right]$ y las curvas estimadas serán 2.38.

2.2.6.3 Método Kalbfleish-Prentice

[22]

$$\hat{S}_0 = \prod_{j=0}^{i-1} \hat{\alpha}_j \quad t_{(i-1)} < t \leq t_i\ , \ \ i = 1, \ldots, k+1$$

donde $\hat{\alpha}_0 \equiv 1$ y $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ se obtiene al solucionar un sistema de k ecuaciones de la forma:

$$\sum_{j \in u^*_{(i)}} \frac{exp\left(\hat{\beta} X_j'\right)}{1 - \hat{\alpha}^{exp\left(\hat{\beta} X_j'\right)}} = \sum_{l \in R(t_{(i)})} \exp\left(\hat{\beta} X_l'\right) \quad \text{para } i = 1, \ldots, k \tag{2.41}$$

Cuando no hay empates la solución viene dada por:

$$\hat{\alpha}_i = \left[1 - \frac{exp\left(\hat{\beta} X_{(i)}'\right)}{\sum_{l \in R(t_{(j)})} exp\left(\hat{\beta} X_l'\right)} \right]^{exp\left(-\hat{\beta} X_{(j)}'\right)} \qquad para\ i = 1, \ldots, k$$

Lo que nos lleva a:

$$\hat{S}_0(t) = \prod_{j=0}^{i-1} \left[1 - \frac{\exp\left(\hat{\beta} X'_{(i)}\right)}{\sum_{l \in R\left(t_{(j)}\right)} \exp\left(\hat{\beta} X'_l\right)} \right]^{\exp\left(-\hat{\beta} X'_{(j)}\right)} \qquad t_{(i-1)} < t \leq t_i\,, \quad i = 1, \ldots, k{+}1 \quad (2.42)$$

Y las curvas estimadas serán 2.38.

2.2.7 EVALUACIÓN DE LA HIPÓTESIS DE RIESGOS PROPORCIONALES

A continuación, explicaremos tres formas de evaluar la hipótesis de riesgos proporcionales, sin embargo, solo entraremos en detalle en una de ellas, el estudio de la bondad de ajuste:

- ❒ **Métodos gráficos**. - El más popular es el uso de los gráficos denominados “log-log”. Si la hipótesis de riesgos proporcionales se cumple estos gráficos deben ser aproximadamente paralelos para las diferentes clases de variables explicativas del modelo por separado. Otro método es el de comparar las curvas de supervivencia “observada” y “esperada” para las diferentes clases de variables explicativas por separado. Éstas curvas deberían ser parecidas, es decir, deberían superponerse si las ponemos en el mismo eje de coordenadas.

- ❒ **Método del uso de variables dependientes del tiempo**. - Cuando utilizamos variables predictoras dependientes del tiempo con el objetivo de evaluar la hipótesis de riesgos proporcionales el modelo se “extiende” con la inclusión del producto, es decir, la interacción, de las covariables con funciones del tiempo, $X \times g$, donde g es una función del tiempo. Por ejemplo, supongamos que queremos evaluar si la variable v (para nuestro ejemplo, única variable explicativa) cumple o no la hipótesis de riesgos proporcionales, podemos tomar $h(t, v) = h_0(t) \exp(\beta v + \delta(v \times t))$ y evaluar a través de una prueba si $\delta = 0$.

- ❒ **Método de Contraste de bondad de ajuste (goodness-of-fit, GOF)**. - La hipótesis de riesgos proporcionales se testea mediante una prueba estadística para cada variable del modelo y el modelo en general. A través del test estadístico tenemos una medida más objetiva que los métodos gráficos y computacionalmente más eficiente

que la inclusión de variables dependientes del tiempo. Entraremos en más detalle en seguida.

2.2.7.1 Método de Contraste de Bondad de Ajuste (GOF)

Se han propuesto diferentes contrastes para estudiar la hipótesis de riesgos proporcionales; uno de los más utilizados es el test propuesto por Harrell y Lee [24].

El método se basa en los denominados residuos de Schoenfeld, los cuales se calculan para cada predictor y para cada individuo no censurado. Si por ejemplo tenemos p predictores en el modelo y n sujetos no censurados, para cada sujeto con tiempo de fallo tendremos p residuos de Schoenfeld. Los residuos de Schoenfeld se definen como:

$$R_{ij} = x_{ij} - \frac{\sum_{l \in R(t_{(i)})} x_{lj} \exp\left(\hat{\beta} X_l'\right)}{\sum_{l \in R(t_{(i)})} \exp\left(\hat{\beta} X_l'\right)} \qquad para\ i = 1, \ldots, n \ \ y \ \ j = 1, \ldots, p$$

Se podría ver los residuos, así calculados, como la diferencia entre la j-ésima covariable del i-ésimo sujeto y la media ponderada de la j-ésima variable de los sujetos a riesgo en el momento de fallo del i-ésimo sujeto, con el valor de riesgo de cada individuo como peso.

Para un predictor en particular se cumplirá la hipótesis de riesgos proporcionales si sus residuos de Schoenfeld no están correlacionados con los tiempos de fallo. Gráficamente, si dibujamos los residuos de Schoenfeld del predictor, éstos serán horizontales si se cumple la hipótesis de proporcionalidad ya que en tal caso los residuos son independientes del tiempo.

Aplicamos este método siguiendo los pasos a continuación:

1. Calcular los residuos de Schoenfeld del modelo de Cox en estudio.
2. Ordenar los tiempos de fallo y crear la variable orden asignando a cada tiempo su número ordinal correspondiente.
3. Calcular la correlación entre los residuos y la variable de orden creada.
4. Realizamos el contraste de hipótesis, donde la hipótesis nula es que la correlación calculada en el paso 3 es cero, es decir, $H_0 : \rho = 0$ para cada predictor por separado.

5. Si se aceptase la hipótesis nula de que la correlación es cero, se cumpliría la hipótesis de riesgos proporcionales para la covariable correspondiente, por lo que nos interesará que los p-valores del contraste sean elevados.

2.2.8 VERIFICACIÓN DE LOS SUPUESTOS DEL MODELO DE COX

Este apartado trataremos los supuestos 2, 3 y 4, ya que el supuesto 1 (La hipótesis de riesgos proporcionales) fue tocada en el apartado anterior. Cada uno de los supuestos se validará mediante el análisis de residuos, por lo que explicaremos cada residuo y su uso.

2.2.8.1 Residuos de Martingala

Los residuos de martingala tienen una distribución sesgada con media cero, son muy asimétricos y con una cola muy larga hacia la derecha, particularmente para datos de supervivencia para un solo evento. Se aplica principalmente para la verificación de la adecuada forma funcional de las covariables continuas (supuesto 2) .

Se los calcula mediante la fórmula:

$$R_{M_i} = \delta_i - R_i \quad parai = 1, \ldots, n \tag{2.43}$$

Donde δ es el indicador de si el sujeto no está censurado, es decir, $\delta_i = 1$ si el sujeto no está censurado y $\delta_i = 0$ si lo está; $R_i = -\log \hat{S}\left(t_i, X_i\right)$ son los denominados residuos de Cox-Snell extendidos para el i-ésimo individuo con tiempo de supervivencia observado t_i y valor de covariables X_i . Si el modelo de Cox es apropiado, el gráfico de R_i y su función de supervivencia estimada de Kaplan-Meier aparecería como una línea recta a 45° . El método de residuos de Cox-Snell es útil para evaluar la bondad del ajuste de un modelo paramétrico.

2.2.8.2 Residuos de Desvíos (Desviances)

Los residuos de desvíos también tienen una media de cero, pero se distribuyen simétricamente alrededor de cero cuando el modelo ajustado es adecuado. Los residuos de desvío son positivos para las personas que sobreviven por un tiempo más corto de lo esperado y

negativos para los que sobreviven por más tiempo. Los residuos de desvío a menudo se usan para evaluar la bondad del ajuste de un modelo de riesgos proporcionales y ajuste del modelo para cada sujeto no censurado, lo que finalmente nos permite detectar observaciones atípicas, por lo que el principal uso de los residuos desvíos es la verificación del supuesto de no influencia de los individuos en la estimación del modelo (supuesto 4).

Los residuos de desvíos se definen como:

$$R_{D_i} = sign\left(R_{M_i}\right)\sqrt{2\left[-R_{M_i} - \delta_i \log\left(\delta_i - R_{M_i}\right)\right]} \qquad para\ i = 1, \ldots, n \tag{2.44}$$

2.2.8.3 Residuos de Puntajes (Scores)

Los residuos de puntajes se calculan para cada individuo y cada covariable. Los “dfbeta” y “dfbetas” (el “dfbeta” estandarizado) se calculan para cada sujeto e indican el cambio aproximado en el vector de coeficientes si el sujeto concreto no estuviese en el modelo. Los residuos de scores se utilizan con la finalidad de verificar la no influencia de los individuos en la estimación de cada coeficiente del modelo (supuesto 3) y para la estimación robusta de la varianza.

Para la j-ésima covariable los residuos de puntajes se definen como:

$$R_{ij}\left(t\right) = \int_0^t \left\{x_{ij}\left(u\right) - \overline{x}_j\left(u\right)\right\} d\hat{M}_i\left(u\right) \qquad para\ i = 1, \ldots, n \tag{2.45}$$

donde

$$\overline{x}_j\left(t\right) = \frac{\sum_{l \in R\left(t_{(i)}\right)} x_{lj} \exp\left(\hat{\beta} X_l'\left(t\right)\right)}{\sum_{l \in R\left(t_{(i)}\right)} \exp\left(\hat{\beta} X_l'\left(t\right)\right)} = \frac{\sum_{i=1}^{n} J_i\left(t\right) x_{ij} \exp\left(\hat{\beta} X_i'\left(t\right)\right)}{\sum_{i=1}^{n} J_i\left(t\right) \exp\left(\hat{\beta} X_i'\left(t\right)\right)} \text{ y}$$

$$\hat{M}_i\left(u\right) = N_i\left(u\right) - \int_0^t J_i\left(s\right) \exp\left(\hat{\beta} X_i'\left(s\right)\right) d\hat{H}_0\left(s\right)$$

Siendo $J_i\left(t\right) = 1$ si el individuo i esta en riesgo al tiempo t y $N_i\left(t\right)$ el indicador del proceso de conteo de si el i-ésimo individuo le ha ocurrido el evento de interés. Estos residuos tienen en cuenta variables dependientes del tiempo y su expresión se particulariza cuando las variables son independientes del tiempo.

2.3 MEDIDAS DE DISCRIMINACIÓN Y PREDICCIÓN DE LOS MODELOS

Para la evaluación de los modelos resultantes usaremos una adaptación de las definiciones de dos medidas de discriminación y precisión de predicción muy conocidas, el área bajo la curva ROC (AUC) y la puntuación de Brier esperada (BS), propuestas por Blanche en 2015 [25] y programadas en lenguaje R[19] en el paquete riskReggression [26]. Debemos subrayar que la segunda medida nombrada, aunque no es tan conocida como la primera, no es nueva, de hecho fue propuesta por Glenn W. Brier en 1950 [27].

Estas adaptaciones de las medidas son básicamente una cuantificación de los indicadores en el tiempo, es decir, son dinámicas o tiempo dependientes.

Para lo que sigue, asumimos una muestra de n sujetos independientes idénticamente distribuidos $\{(\widetilde{T}_i,\ \Delta_i,\ \widetilde{\eta}_i,\ \pi_i(\cdot,\cdot)),\ i = 1, ..., n\}$, donde $\pi_i(\cdot,\cdot)$ denota un proceso de predicción específico del i-ésimo sujeto computable para todos los tiempos de referencia s y en el horizonte de predicción t. Debemos tener en cuenta que asumir una muestra i.i.d. implica que el modelo conjunto utilizado para calcular $\pi_i(s,t)$ a partir de las covariables del sujeto i se ha ajustado a un conjunto de datos independiente (En el capítulo de resultados, más adelante, la llamaremos base de desarrollo y a la data sobre la que calculamos las medidas será la base de prueba). Sin pérdida de generalidad, establecemos $\pi_i(s,t) = 0$ para todos los sujetos i que ya no están en riesgo en s, y nos enfocamos en la predicción del evento $\eta = 1$ (evento de interés).

2.3.1 AUC TIEMPO DEPENDIENTE

En la práctica, nos gustaría una herramienta de predicción que pronostique un mayor riesgo de que les ocurra el evento de interés a los sujetos que verdaderamente tienen más probabilidades de experimentarlo que para los sujetos que tienen menos probabilidades de que les suceda. Este es el concepto de precisión predictiva en términos de discriminación, para el cual el AUC es una medida significativamente útil.

Formalmente, al combinar la definición de curva ROC para riesgos en competencia [28][29] y la de predicción dinámica [30], Blanche propone la siguiente definición del AUC tiempo

dependiente (AUC dinámico), en el instante s para un horizonte de predicción t:

$$AUC(s,t) = \mathcal{P}\left(\pi_i(s,t) > \pi_j(s,t) \mid D_i(s,t) = 1, D_j(s,t) = 0, T_i > s, T_j > s\right)$$

donde $D_i(s,t) = I(s < T_i \leq s + t, \eta_i = 1)$. Con esta notación, para cualquier sujeto i en riesgo en el instante s, $D_i(s,t) = 1$ cuando el sujeto i experimenta el evento de interés dentro del intervalo de tiempo $(s, s+t]$, y $D_i(s,t) = 0$ cuando el sujeto i no experimenta el evento de interés hasta $s + t$. Dentro de la terminología de la metodología ROC, en el instante s y el horizonte de predicción t, el sujeto i en riesgo en el momento s se define como un caso cuando $D_i(s,t) = 1$ y un control cuando $D_i(s,t) = 0$[29].

2.3.2 BRIER TIEMPO DEPENDIENTE ESPERADO

Al combinar las definiciones del puntaje Brier esperado para riesgos en competencia [31] y para la predicción dinámica [32], Blanche propone la siguiente definición para el puntaje Brier tiempo dependiente (dinámico) esperado:

$$BS(s,t) = E\left[\left(D(s,t) - \pi(s,t)\right)^2 \mid T > s\right],$$

2.3.3 CONTRASTES ENTRE AUC y BS

Tanto AUC como BS son interesantes y se complementan entre sí para medir qué tan buena es una herramienta de predicción dinámica $\pi(s,t)$, donde $\pi(s,t)$ depende de algunos parámetros estimados $\hat{\xi}$ y las covariables de referencia **X**. El AUC es particularmente conveniente para fines de comunicación, ya que tiene una interpretación simple como el índice de concordancia, no depende de la incidencia acumulada del evento de interés $\mathcal{P}(s < T \leq s + t, \eta = 1 | T > s)$, y por lo tanto tiene una escala fácilmente comprensible. Por el contrario, BS tiene la ventaja de ser una medida de precisión predictiva más completa, ya que cuantifica tanto la calibración como la discriminación.

2.3.4 ESTIMACIÓN DE AUC Y BS TIEMPO DEPENDIENTES

En presencia de datos censurados inducidos por la pérdida durante el seguimiento, para todos los sujetos censurados dentro de $(s, s+t)$, el indicador $D_i(s,t)$ no se puede calcular y, por lo tanto, se desconoce. Para superar esta "falta de datos", Blanche menciona la técnica de probabilidad inversa de censura ponderada (IPCW, por su siglas en inglés) y menciona algunos ejemplos en los que se ha aplicado en varios entornos estrechamente relacionados a su trabajo [33][29][30].

Por esta razón, Blanche propone usar los estimadores IPCW:

$$\widehat{AUC}(s,t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I_{(\pi_i(s,t)>\pi_j(s,t))}\widetilde{D}_i(s,t)\left(1-\widetilde{D}_j(s,t)\right)\widehat{W}_i(s,t)\widehat{W}_j(s,t)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{D}_i(s,t)\left(1-\widetilde{D}_j(s,t)\right)\widehat{W}_i(s,t)\widehat{W}_j(s,t)},$$

y

$$\widehat{BS}(s,t) = \frac{1}{n\widehat{S}_{\widetilde{T}}(s)}\sum_{i=1}^{n}\widehat{W}_i(s,t)\left(\widetilde{D}_i(s,t)-\pi_i(s,t)\right)^2,$$

donde $\widehat{S}_{\widetilde{T}}(s) = (1/n)\sum_{i=1}^{n} I_{\widetilde{T}_i>s}$ estima la probabilidad de observar un sujeto en riesgo en s. El indicador $\widetilde{D}_i(s,t) = I(s < \widetilde{T}_i \leq s+t, \widetilde{\eta}_i = 1)$ es igual a 1 cuando se sabe que el sujeto i ha experimentado el evento de interés dentro de $(s, s+t]$, y es igual a 0 en caso contrario. Para tener en cuenta la censura, los pesos se definen como

$$\widehat{W}_i(s,t) = \frac{I_{\widetilde{T}_i>s+t}}{\widehat{G}(s+t|s)} + \frac{I_{s<\widetilde{T}_i\leq s+t}\Delta_i}{\widehat{G}(\widetilde{T}_i|s)}$$

donde $\widehat{G}(u)$ es el estimador de Kaplan-Meier de la función de supervivencia del tiempo de censura en u, es decir, $\mathcal{P}(C > u)$, y $\forall u > s$, $\widehat{G}(u|s) = \widehat{G}(u)/\widehat{G}(s)$ estima la probabilidad condicional de no ser censurado en el instante u condicionado a no estar censurado en el instante s.

Es importante destacar que estos estimadores son independientes del modelo en el sentido de que no existe ningún supuesto sobre la exactitud de la especificación del modelo conjunto utilizado para calcular $\pi_i(s,t),\ i = 1, ..., n$.

2.4 ÁRBOLES DE INFERENCIA CONDICIONAL

La mayoría de los algoritmos de partición recursiva son casos especiales de un algoritmo simple de dos etapas:

1. Particionar las observaciones mediante divisiones univariadas de forma recursiva.

2. Ajustar un modelo constante en cada celda de la partición resultante.

Las implementaciones más populares de tales algoritmos son '**CART**'y '**C4.5**'. Al igual que **AID**, ambos realizan una búsqueda exhaustiva sobre todas las posibles divisiones maximizando una medida de información de la impureza del nodo seleccionando la covariable que muestra la mejor división. Este enfoque tiene dos problemas fundamentales: sobre-entrenamiento y un sesgo de selección hacia covariables con muchas divisiones posibles. Con respecto al problema de sobre-entrenamiento, Mingers [34] señala que el algoritmo

"... no tiene un concepto de significación estadística, por lo que no puede distinguir entre una mejora significativa y una insignificante en la medida de información."(Traducido del inglés).

Hothorn y colaboradores [35] hacen una descripción completa de los fundamentos metodológicos de los árboles de inferencia condicionales mediante los cuales puede enfrentarse los requerimientos de White y Liu (1994) quienes exigen

"... un enfoque estadístico (de la partición recursiva) que tenga en cuenta las propiedades distributivas de las medidas".(Traducido del inglés).

Un marco unificado que incorpora la partición binaria recursiva en la teoría bien definida de las pruebas de permutación desarrollada por Strasser y Weber [36] es presentado en el trabajo de Hothorn y colaboradores [35]. La distribución condicional de estadísticos que miden la asociación entre respuestas y covariables es la base para una selección insesgada entre covariables medidas en diferentes escalas. Además, se aplican múltiples procedimientos de prueba para determinar si no se puede establecer una asociación significativa entre cualquiera de las covariables y la respuesta y si es necesario detener la recursividad.

2.4.1 PARTICIÓN BINARIA RECURSIVA

Hothorn y colaboradores [35] se centran en modelos de regresión que describen la distribución condicional de una variable de respuesta **Y** dado el estado de m covariables mediante partición recursiva estructurada en árbol. La respuesta **Y** de algún espacio muestral $\mathcal{Y}$ también puede ser multivariante. El vector covariable m-dimensional $\mathbf{X} = (X_1, ..., X_m)$ se toma de un espacio muestral $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$. Tanto la variable de respuesta como las covariables pueden medirse a escalas arbitrarias. Suponemos que la distribución condicional $D(\mathbf{Y}|\mathbf{X})$ de la respuesta **Y** dadas las covariables **X** depende de una función f de las covariables

$$D(\mathbf{Y}|\mathbf{X}) = D(\mathbf{Y}|X_1, ..., X_m) = D(\mathbf{Y}|f(X_1, ..., X_m)),$$

donde nos restringimos a relaciones de regresión basadas en particiones, es decir, r celdas disjuntas $B_1, ..., B_r$ dividiendo el espacio de covariables $\mathcal{X} = \cup_{k=1}^{r} B_k$. Se ajustará un modelo de la relación de regresión en función de una muestra de aprendizaje $\mathcal{L}_n$, es decir, una muestra aleatoria de n observaciones independientes e idénticamente distribuidas, posiblemente con algunas covariables X_{ji} faltantes,

$$\mathcal{L}_n = \{(\mathbf{Y}_i, X_{1i}, ..., X_{mi}); i = 1, ..., n\}$$

Se puede formular un algoritmo genérico para el particionamiento binario recursivo para una muestra de aprendizaje dada $\mathcal{L}_n$ utilizando pesos y casos con valores enteros no negativos $\mathbf{w} = (w_1, ..., w_n)$. Cada nodo de un árbol está representado por un vector de pesos y casos que tiene elementos distintos de cero cuando las observaciones correspondientes son elementos del nodo y son cero en caso contrario. El siguiente algoritmo implementa particiones binarias recursivas:

1. Para los pesos y casos **w**, se prueba la hipótesis nula global de independencia entre cualquiera de las m covariables y la respuesta. El algoritmo de detiene si esta hipótesis no puede ser rechazada. De lo contrario, selecciona la j^*-ésima covariable X_{j^*} con la asociación más fuerte a Y.

2. Elija un conjunto $A^* \subset \mathcal{X}_{j^*}$ para dividir X_j^* en dos conjuntos separados A^* y $X_{j^*} \setminus A^*$. Los pesos y casos $\mathbf{w}_{izq}$ y $\mathbf{w}_{der}$ (izq: izquierda y der: derecha) determinan los dos subgrupos con $w_{izq,i} = w_i I(X_{j^*i} \in A^*)$ y $w_{der,i} = w_i I(X_{j^*i} \notin A^*)$ para todo $i = 1, ..., n$ ($I(\cdot)$ denota la función Indicador).

3. Repita de forma recursiva los pasos 1 y 2 con pesos y casos modificados w_{izq} y w_{der}, respectivamente.

Hothorn y colaboradores comentan que la separación de la selección de variables y el procedimiento de división en los pasos 1 y 2 del algoritmo es la clave para la construcción de estructuras de árbol interpretables que no sufran una tendencia sistemática hacia covariables con muchas divisiones posibles o muchos valores perdidos. Además, se puede implementar un criterio de parada intuitivo y estadísticamente motivado: nos detenemos cuando la hipótesis nula global de independencia entre la respuesta y cualquiera de las m covariables no puede ser rechazada en un nivel nominal predeterminado α. El algoritmo induce una partición $B1, ..., Br$ del espacio covariable $\mathcal{X}$, donde cada celda $B \in B_1, ..., B_r$ está asociado con un vector de pesos y casos.

2.4.2 PARTICIONAMIENTO RECURSIVO POR INFERENCIA CONDICIONAL

La parte principal de esta sección se enfocará en el Paso 1 del algoritmo genérico. Las pruebas unificadas de independencia se construyen mediante la distribución condicional de estadísticos lineales en el marco de pruebas de permutación desarrollado por Strasser y Weber [36]. La determinación de la mejor división binaria en una covariable seleccionada y el manejo de los valores perdidos se realiza también en base a estadísticas lineales estandarizadas dentro del mismo marco.

2.4.2.1 Criterios de Selección Variables y Parada

En el paso 1 del algoritmo genérico dado en la página 2.4.1 nos enfrentamos a un problema de independencia. Necesitamos decidir si hay alguna información sobre la variable de respuesta cubierta por alguna de las m covariables. En cada nodo identificado por pesos y casos **w**, la hipótesis global de independencia se formula en términos de las m hipótesis parciales $H_0^j : D(\mathbf{Y}|X_j) = D(\mathbf{Y})$ con hipótesis nula global $H_0 = \cap_{j=1}^{m} H_0^j$. Cuando no podemos rechazar H_0 en un nivel preespecificado α, detenemos la recursividad. Si la hipótesis global puede ser rechazada, medimos la asociación entre **Y** y cada una de las covariables X_j, $j = 1, ..., m$, mediante estadísticos de prueba o p-valores que indican la desviación de

las hipótesis parciales H_0^j.

Por conveniencia de notación y sin pérdida de generalidad, asumimos que los pesos y casos w_i son cero o uno. El grupo simétrico de todas las permutaciones de los elementos de $(1, ..., n)$ con los correspondientes pesos y casos $w_i = 1$ se denota por $S(\mathcal{L}_n, \mathbf{w})$. Medimos la asociación entre **Y** y X_j, $j = 1, ..., m$, por estadísticos lineales de la forma

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = vec\left(\sum_{i=1}^{n} w_i g_i(X_{ji}) h(\mathbf{Y}_i, (\mathbf{Y}_1, ..., \mathbf{Y}_n))^T\right) \in R^{p_j q} \tag{2.46}$$

donde $g_j : \mathcal{X}_j \to R^{p_j}$ es una transformación no aleatoria de la covariable $\mathbf{X}_j$. La función de influencia $h : \mathcal{Y} \times \mathcal{Y}^n \to R^q$ depende de las respuestas $(\mathbf{Y}_1, ..., \mathbf{Y}_n)$ de forma simétrica por permutación. Hothorn [35] explica cómo elegir g_j y h en diferentes entornos prácticos. Una matriz $p_j \times q$ se convierte en un vector de columna $p_j q$ mediante una combinación de columnas utilizando el operador "vec".

La distribución de $\mathbf{T}_j(\mathcal{L}_n, w)$ bajo H_0^j depende de la distribución conjunta de **Y** y X_j, que se desconoce en casi todas las circunstancias prácticas. Al menos bajo la hipótesis nula, uno puede deshacerse de esta dependencia fijando las covariables y condicionando todas las posibles permutaciones de las respuestas. Este principio conduce a procedimientos de prueba conocidos como *pruebas de permutación*. La esperanza condicional $\mu_j \in R^{p_j q}$ y la covarianza $\Sigma_j \in R^{p_j q \times p_j q}$ de $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ bajo H_0 dadas todas las permutaciones $\sigma \in S(\mathcal{L}_n, \mathbf{w})$ de las respuestas fueron derivadas por Strasser y Weber [36]:

$$\begin{aligned}
\mu_j &= E(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) | S(\mathcal{L}_n, \mathbf{w})) = vec\left(\left(\sum_{i=1}^{n} w_i g_i(X_{ji})\right) E(h | S(\mathcal{L}_n, \mathbf{w}))^T\right), \\
\Sigma_j &= V(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) | S(\mathcal{L}_n, \mathbf{w})) \\
&= \frac{\mathbf{w}.}{\mathbf{w}. - 1} V(h | S(\mathcal{L}_n, \mathbf{w})) \otimes \left(\sum_i w_i g_j(X_{ji}) \otimes w_i g_j(X_{ji})^T\right) \\
&\quad - \frac{1}{\mathbf{w}. - 1} V(h | S(\mathcal{L}_n, \mathbf{w})) \otimes \left(\sum_i w_i g_j(X_{ji})\right) \otimes \left(\sum_i w_i g_j(X_{ji})\right)^T,
\end{aligned} \tag{2.47}$$

donde $\mathbf{w}. = \sum_{i=1}^{n} w_i$ denota la suma de los pesos y casos, $\otimes$ es el producto de Kronecker y la esperanza condicional de la función de influencia es

$$E(h | S(\mathcal{L}_n, \mathbf{w})) = \mathbf{w}.^{-}1 \sum_i w_i h(\mathbf{Y}_i, (\mathbf{Y}_1, ..., \mathbf{Y}_n)), \in R^q$$

Con su matriz de covarianza $q \times q$ correspondiente

$$V(h|S(\mathcal{L}_n, \mathbf{w})) = \mathbf{w}.^{-}1 \sum_i w_i (h(\mathbf{Y}_i, (\mathbf{Y}_1, ..., \mathbf{Y}_n)) - E(h|S(\mathcal{L}_n, \mathbf{w})))$$
$$(h(\mathbf{Y}_i, (\mathbf{Y}_1, ..., \mathbf{Y}_n)) - E(h|S(\mathcal{L}_n, \mathbf{w})))^T.$$

Teniendo a mano la esperanza condicional y la covarianza, podemos estandarizar un estadístico lineal $\mathbf{T} \in R^{pq}$ de la forma 2.46 para algún $p \in p_1, ..., p_m$. Los estadísticos de prueba univariados c que envían un estadístico lineal multivariante observado $\mathbf{t} \in R^{pq}$ en la recta real pueden ser de forma arbitraria. Una elección obvia es el máximo de los valores absolutos del estadístico lineal estandarizado

$$c_{max}(\mathbf{t}, \mu, \Sigma) = \max_{k=1,...,pq} \left| \frac{(\mathbf{t} - \mu)k}{\sqrt{(\Sigma)kk}} \right|$$

utilizando la esperanza condicional μ y la matriz de covarianza Σ. La aplicación de una forma cuadrática $c_{quad}(\mathbf{t}, \mu, \Sigma) = (\mathbf{t} - \mu)\Sigma^+(\mathbf{t} - \mu)^T$ es una alternativa, aunque computacionalmente más costosa porque está involucrada la inversa de Moore-Penrose Σ^+ de Σ. Es importante señalar que los estadísticos de prueba $c(\mathbf{t}j, \mu_j, \Sigma_j)$, $j = 1, ..., m$, no se pueden comparar directamente de manera insesgada a menos que todas las covariables se midan en la misma escala, es decir, $p_1 = p_j$, $j = 2, ..., m$. Para permitir una selección de variable insesgada, necesitamos cambiar a la escala del P-valor porque los P-valores para la distribución condicional de las estadísticos de prueba $c(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}), \mu_j, \Sigma_j)$ se pueden comparar directamente entre las covariables medidas a diferentes escalas. En el paso 1 del algoritmo genérico seleccionamos la covariable con P-valor mínimo, es decir, la covariable X_{j^*} con $j^* = \underset{j=1,...,m}{\operatorname{argmin}} P_j$, donde

$$P_j = \mathcal{P}_{H_0^j}(c(\mathbf{T}_j, \mu_j, \Sigma_j) \geq c(\mathbf{t}_j, \mu_j, \Sigma_j)|S(mathcalL_n, \mathbf{w}))$$

es el P-valor de la prueba condicional para H_0^j.

Hasta ahora, solo hemos abordado la prueba de cada hipótesis parcial H_0^j, que es suficiente para una selección de variable insesgada. Se puede construir una prueba global para H_0 requerida en el paso 1 mediante una agregación de las transformaciones $g_j, j = 1, ..., m$, es decir, usando un estadístico lineal de la forma

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = vec\left(\sum_{i=1}^{n} w_i \left(g_1(X_{1i})^T, ..., g_m(X_{mi})^T \right)^T h\left(\mathbf{Y}_i, (\mathbf{Y}_1, ..., \mathbf{Y}_n) \right)^T \right).$$

Sin embargo, este enfoque es menos atractivo para muestras de desarrollo con valores perdidos. Los enfoques de aplicación universal son múltiples procedimientos de prueba basados en $P_1, ..., P_m$. Los P-valores simples ajustados por Bonferroni o un enfoque de remuestreo del mínimo P-valor son solo ejemplos y nos referimos a la literatura de pruebas múltiples (por ejemplo, Westfall y Young [37]) para métodos más avanzados. Se rechaza H_0 cuando el mínimo de los P-valores ajustados es menor que un nivel nominal α preespecificado o, de lo contrario, se detiene el algoritmo. En este sentido, α puede verse como un parámetro único que determina el tamaño de los árboles resultantes.

2.4.2.2 Criterios de división

Una vez que hemos seleccionado una covariable en el paso 1 del algoritmo, la división en sí puede establecerse mediante cualquier criterio de división, incluidos los establecidos por Breiman y colaboradores [38] o Shih [39]. En lugar de simples divisiones binarias, también se pueden implementar divisiones de múltiples vías, por ejemplo, utilizando el trabajo de O'Brien [40]. Sin embargo, la mayoría de los criterios de división no son aplicables a las variables de respuesta medidas a escalas arbitrarias y, por lo tanto, utilizamos el marco de prueba de permutación descrito anteriormente para encontrar la división binaria óptima en una covariable seleccionada X_{j^*} en el paso 2 del algoritmo genérico. La bondad de una división se evalúa mediante estadísticos lineales de dos muestras, que son casos especiales del estadístico lineal 2.46. Para todos los posibles subconjuntos A del espacio muestral $\mathcal{X}_{j^*}$ el estadístico lineal

$$\mathbf{T}_{j^*}^A(\mathcal{L}_n, \mathbf{w}) = vec\left(\sum_{i=1}^{n} w_i I(X_{j^*i} \in A) h(\mathbf{Y}_i, (\mathbf{Y}_1, ..., \mathbf{Y}_n))^T\right) \in R^q$$

induce un estadístico de dos muestras que mide la discrepancia entre las muestras $\{Y_i | w_i > 0 y X_{ji} \in A;\ i = 1, ..., n\}$ y $\{Y_i | w_i > 0 y X_{ji} \notin A;\ i = 1, ..., n\}$. La esperanza condicional $\mu_{j^*}^A$ y la covarianza $\Sigma_{j^*}^A$ pueden calcularse mediante 2.47. Se establece la división A^* con un estadístico de prueba maximizado sobre todos los posibles subconjuntos A:

$$A^* = \underset{A}{\operatorname{argmax}}\, c(\mathbf{t}_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A). \tag{2.48}$$

Tenga en cuenta que no es necesario calcular la distribución de $c(\mathbf{t}_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A)$ en el Paso 2. Para evitar divisiones patológicas, se puede restringir el número de posibles subconjuntos

que se evalúan, por ejemplo, mediante la introducción de restricciones sobre el tamaño de la muestra o la suma de las ponderaciones de los casos en cada uno de los dos grupos de observaciones inducidas por una posible división.

2.4.2.3 Valores perdidos y divisiones sustitutas

Si falta una observación X_{ji} en la covariable X_j, establecemos el peso y caso correspondiente w_i en cero para el cálculo de $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ y, si quisiéramos dividir en X_j, también en $\mathbf{T}_j^A(\mathcal{L}_n, \mathbf{w})$. Una vez que se ha implementado una división A^* en X_j, se pueden establecer divisiones sustitutas buscando una división que conduzca aproximadamente a la misma división de las observaciones que la división original. Uno simplemente reemplaza la variable de respuesta original por una variable binaria $I(X_{ji} \in A^*)$ que codifica la división y procede como se describe en la parte anterior.

3 METODOLOGÍA

Antes de atender los objetivos específicos del presente trabajo de titulación se realizará una recopilación bibliográfica que pretende describir teórica y formalmente el análisis de supervivencia, así como el modelo de riesgo proporcional o modelo de Cox. De la misma manera, se describirá formalmente la distancia de Mahalanobis, la misma que luego nos permitirá eliminar casos atípicos, y los árboles de decisión, los cuales nos ayudarán a discretizar algunos predictores[1] y también crear nuevas variables.

Luego, se abordará cuatro temas importantes que permitirán cumplir con el primer objetivo específico. El primer tema que permitirá tener una visión completa de la medicina prepagada en el Ecuador y compararla con la situación en otros países tanto de América latina como del mundo. Un segundo tema con el que se sustituirá el uso del modelo de manera empírica, mostrando resultados obtenidos en otras ramas de negocio. En tercer tema que trataremos será esencial para darnos un vistazo de las consecuencias de la caída de cartera en las empresas de medicina prepagada, ya que la pérdida de ingresos es solo la punta del iceberg. Por último, trataremos como otras empresas de medicina prepagada alrededor del mundo controlan su riesgo de deserción de clientes.

Posteriormente, se describirá a los clientes objetivo de este análisis con el fin de evaluar la situación actual de la empresa patrocinadora con respecto al riesgo de caída de cartera en la población seleccionada.

Seguidamente, se definiré de manera estricta cual es el evento de interés y ya armados con una definición clara se seguirá el proceso metodológico definido a continuación:

1. Descripción de los sujetos de estudio

[1] Esta es una técnica muy usual en los modelos de Cox y en los apartados oportunos explicaremos las razones para hacerlo y discutiremos las propuestas de algunos autores.

2. Definición del evento de riesgo que caracteriza la no supervivencia (deserción del cliente)
3. Selección de la Ventana de muestreo
4. Análisis descriptivo de base bruta
5. Decisión de partición en dos modelos
6. Muestreo para partición en bases de desarrollo y prueba
7. Análisis descriptivo y depuración de las bases de datos por cada población
8. Creación de variables discretas a partir de variables continuas
9. Estimación del modelo de Cox
10. Validación de supuestos de los modelos de Cox
11. Validación en base de prueba
12. Creación de perfiles de riesgo

Terminado este proceso se habrá cumplido con el segundo objetivo.

Finalmente, en atención al último objetivo específico se automatizará el análisis de supervivencia a través de la programación de una función en el software R, con la que se pretende la reproducción y parametrización del análisis. De esta manera, se podrá entregar información del modelo para las áreas pertinentes. Debido a lo reservado de los datos y del modelo, la mayor parte de ese script será cambiado para que no se pueda reproducir por una fuente externa.

3.1 DETALLE DE ACTIVIDADES RELEVANTES

A continuación, se detallará cada una de las actividades que se ejecutó para poder llegar a un modelo adecuado. Las técnicas serán descritas en cada uno de los numerales, si se lo considera necesario, o de lo contrario será adjuntado en los anexos, en cualquier caso, el lector tendrá una fuerte explicación de todo lo realizado.

3.1.1 DESCRIPCIÓN DE LOS CLIENTES OBJETIVOS DEL ANÁLISIS DE SUPERVIVENCIA

En esta sección se describirá, valga la redundancia, a los clientes de la línea individual, los cuales fueron seleccionados para este trabajo. Se responderán preguntas como ¿Cuál es la edad promedio de titulares, cónyuge e hijos? ¿Cuántos afiliados tiene cada contrato? mientras que otras preguntas como lo que se refiere a prima promedio y planes con mayor adquisición no se toparan debido a lo delicado de la información.

3.1.2 DEFINICIÓN DEL EVENTO DE RIESGO QUE CARACTERIZA LA NO SUPERVIVENCIA (DESERCIÓN DEL CLIENTE)

Esta definición es muy impórtate debido a que si es errónea dará resultados peligrosos, por lo que se explicará a profundidad cuál es la definición por la que se optó y porque razón no se incluyó algún que otro criterio que parecería relevante. La deserción del cliente puede ser una definición obvia pero no lo es del todo, como ya veremos.

3.1.3 SELECCIÓN DE LA VENTANA DE MUESTREO

Se presentará los criterios para la elección de la ventana de muestreo y que contratos fueron seleccionados, para el análisis.

3.1.4 ANÁLISIS DESCRIPTIVO Y DEPURACIÓN DE LA BASE BRUTA

Se presentará algunos de los problemas de la base bruta y como se depuró algunos problemas (cuando esto fue posible) o porqué razón se optó por eliminar ciertos sujetos del estudio o inclusive la variable.

3.1.5 DECISIÓN DE PARTICIÓN EN DOS MODELOS

El estimador de la función de supervivencia, propuesto por Kaplan y Meier (1958), se usará para estimar las funciones de supervivencia de dos grupos de sujetos de estudio que son los contratos con más de un año de afiliación a la compañía y los que tienen un año o menos, de esta manera veremos gráficamente las diferencia entre estos. Debido a que se cree que tienen funciones de riesgo diferentes, usaremos la prueba log-rank para verificar nuestra hipótesis, ésta es quizás la prueba más popular para probar la igualdad (o desigualdad, en nuestro caso) de las funciones de riesgo.

3.1.6 MUESTREO PARA PARTICIÓN EN BASES DE DESARROLLO Y PRUEBA

Se pueden usar muchos métodos de validación para los modelos estadísticos, para este trabajo se seleccionó la partición en bases de desarrollo y prueba. Para hacer esto se hará un muestreo aleatorio estratificado. Se usará 2/3 de los datos para entrenamiento y 1/3 para pruebas.

3.1.7 ANÁLISIS DESCRIPTIVO Y DEPURACIÓN DE LAS BASES DE DATOS POR CADA POBLACIÓN

Se presentará una descripción de cada una de las variables, algunas de las cuales no seguirán en el proceso debido a los valores perdidos.

Se hará un análisis de correlación de variables, mediante el estadístico ρ (rho) de Spearman, para verificar que las variables estén correlacionadas con el evento de interés. Y que además las variables seleccionadas no tengan una correlación fuerte entre ellas y se eliminarán algunas variables más.

3.1.8 CREACIÓN DE VARIABLES DISCRETAS A PARTIR DE VARIABLES CONTINUAS

Mediante árboles de decisión se discretizarán las variables para crear lo que llamaremos vectores. La creación de estas variables discretizadas puede ser un proceso bastante agotador debido a la gran cantidad de información que debe procesarse, por lo que es muy posible que se creen vectores muy parecidos o inclusive los mismos, por esta razón se usará la correlación de Spearman para poder evitar que dos vectores correlacionados ingresen al modelo.

3.1.9 ESTIMACIÓN DEL MODELO DE COX

Usaremos la función coxph para estimar los coeficientes de las covariables de los modelos de Cox para cada uno de los grupos, así como la verificación del supuesto de riesgos proporcionales.

En este punto se dará un panorama más amplio sobre lo que nos dicen los resultados y cómo aplicarlo en el día a día de la empresa y por supuesto dónde se puede mejorar y cómo hacerlo de manera eficaz.

3.1.10 VERIFICACIÓN DE LOS SUPUESTOS DEL MODELO DE COX

Este punto se verificará que los modelos cumplan con los supuestos requeridos para el modelo de Cox de manera gráfica. No se incluye la verificación de la hipótesis de riesgos proporcionales ya que se lo habrá hecho en el punto anterior. Los residuos de la martingala, de puntajes y de desvíos se usarán en este punto.

3.1.11 VALIDACIÓN EN BASE DE PRUEBAS

A través del AUC tiempo dependiente y del Brier tiempo dependiente aplicados a las bases de prueba. verificaremos que los modelos se pueden usar para el propósito que fueron creados.

3.1.12 CREACIÓN DE PERFILES DE RIESGO

Con el modelo final tendremos suficiente información para crear perfiles de riesgo e identificar a los clientes con mayor riesgo de deserción, con el objetivo de que la empresa pueda tomar medidas correctivas que ayuden a evitar la caída de cartera. No obstante, si la capacidad predictiva del modelo no es buena entonces no será posible crear estos perfiles ya que es indispensable.

4 RESULTADOS Y DISCUSIÓN

En este capítulo mostraremos todos los resultados, valga la redundancia, fruto de la investigación realizada. En primer lugar, pondremos en contexto al lector sobre la situación de la medicina prepagada que en estos últimos años cambió radicalmente en el país. Luego, veremos algunas experiencias de la aplicación del análisis de supervivencia y el modelo de riesgos proporcionales en el control de la deserción de clientes, no necesariamente en el mismo sector. Seguidamente, se expondrá cómo las empresas de medicina prepagada han atacado este problema no solo localmente sino también en Latinoamérica y el mundo. Finalmente, el lector podrá apreciar los resultados obtenidos de la aplicación del modelo de Cox en la deserción de clientes de la empresa Humana S.A. conclusiones, comentarios y observaciones. Desde este momento se usará un lenguaje más técnico propio del sector de la medicina prepagada, por lo que se creó un diccionario de términos (ver anexo 7.3) que ayudará al lector a una adecuada interpretación del trabajo presentado.

4.1 SITUACIÓN ACTUAL DE LA MEDICINA PREPAGA EN EL PAÍS Y EL MUNDO

En esta sección estudiaremos algunas restricciones de tipo jurídico que nuestro modelo debe tomar en cuenta para que sea completamente legal implantarlo y usarlo dentro del ejercicio normal de la empresa de medicina prepagada que patrocina este estudio. Además, se discutirá la situación económica de las empresas que brindan servicios integrales de salud prepagada y algunas de las razones por las que el modelo de deserción se vuelve mucho más urgente frente a proyecciones conservadoras que estiman que la inversión en salud para el año 2019 será baja y, en ciertos casos, nula.

4.1.1 MARCO LEGAL DE LAS EMPRESAS DE MEDICINA PREPAGADA EN EL ECUADOR

El 17 de octubre de 2016 se publicó en el registro oficial la ley orgánica que regula a las compañías que financien servicios de atención integral de salud prepagada y a las de seguros que oferten cobertura de seguros de asistencia médica que tiene como objetivo "... normar la constitución y funcionamiento de las compañías que financien servicios de atención integral de salud prepagada; regular, vigilar y controlar la prestación de dichos servicios para garantizar el ejercicio pleno de los derechos de los usuarios; fijar las facultades y atribuciones para establecer y aprobar el contenido de los planes y contratos de atención integral de salud prepagada y de seguros en materia de asistencia médica; así como determinar la competencia para la aplicación del régimen sancionador y la solución de controversias.", según reza el artículo 1 de la misma; esta ley cambió todo el panorama de las empresas de este ramo. La norma antes mencionada incluye alrededor de 46 beneficios para los usuarios de planes de medicina prepagada, los mismos que se centran en tres principios:

1. Regulación por parte del Estado.
2. Precautelar los derechos de los usuarios.
3. Sostenibilidad de las empresas privadas.

Dentro de los beneficios más importantes para los usuarios, y que generaron grandes cambios en la manera que se habían estado manejando las compañías que ofrecen servicios de medicina integral prepagada, están:

1. **La no discriminación**.- Las compañías de medicina prepaga están prohibidas de negarse a celebrar o renovar un contrato en razón de etnia, lugar de nacimiento, edad, sexo, identidad de género, identidad cultural, estado civil, idioma, religión, ideología, filiación política, pasado judicial, condición socio-económica, condición migratoria, orientación sexual, estado de salud, portar VIH, discapacidad, diferencia física; ni por cualquier otra distinción, personal o colectiva, temporal o permanente.
2. **Las carencias de atención de maternidad**.- Estos períodos de carencias no pueden ser superiores a 60 días y si la persona quedase embarazada dentro del período de carencia tendrá derecho a la atención prenatal.

3. **Preexistencias**.- Las compañías deberán cubrir enfermedades preexistentes por un monto anual de 20 salarios básicos unificados del trabajador en general; esta cobertura se aplicara una vez cumplido el período de carencia de 24 meses corridos a partir de la fecha de suscripción del contrato.

4. **Medicina preventiva y Tarifa cero**.- La ley dice que los usuarios podrán acceder a servicios de promoción y prevención. Además, el artículo 30, numeral 3 dice que los afiliados o asegurados deberán recibir prestaciones de prevención primaria con cargo a la tarifa contratada y que estas prestaciones serán determinadas y reguladas por la Autoridad Sanitaria Nacional. En todo caso la ley no es muy clara al respecto por lo que es preferible dejar este punto de lado.

5. **Incremento de Primas**.- En el artículo 29 de la ley dice que la prima que se fije por el plan contratado por el afiliado tienen que estar conforme a los estudios actuariales y notas técnicas aprobadas por los entes de control, es decir, técnicamente no prohíbe el incremento de primas pero si pone un tope a los mismo.

6. **Modificar o eliminar contratos**.- Las empresas no podrán dar por terminado contratos a menos que el usuario incumpla con sus obligaciones o no pague las contraprestaciones económicas en los plazos que se establecieron. Además, no se podrá modificar el precio mientras transcurra el plazo estipulado en el contrato.

7. **Cobertura al adulto mayor**.- Las compañías no podrán modificar, disminuir o restringir coberturas a los afiliados o beneficiarios por el hecho de que sean de edad avanzada o estén próximos a serlo.

8. **Atención ambulatoria para diálisis y hemodiálisis**.- La normativa actual requiere que se incluya en los planes de medicina prepagada la prestación de servicios ambulatorios de diálisis y hemodiálisis.

9. **Enfermedades oncológicas con cobertura integral**.- previo a la ley se ofrecían planes adicionales para cubrir este tipo de enfermedades que no estaban incluidas en los planes normales. Hoy en día los pacientes oncológicos deben ser atendidos de manera integral, incluyendo también cirugías reconstructivas y rehabilitación todo acorde al plan adquirido.

10. **Trasplantes**.- La cobertura de trasplantes está contemplada en la nueva ley y no solo para el paciente sino también para el donante. Se incluye los procesos de pretras-

plante, trasplante y post trasplante además de cubrir complicaciones médicas y otras atenciones necesarias para el paciente y el donante.

11. **La obligación de las empresas de pagar por servicios de salud prestados a sus afiliados en instituciones de la Red Pública Integral de Salud**.- Este es sin duda el punto más controversial de la ley, ya que obliga a las empresas de medicina prepagada a remunerar a las instituciones de salud del Estado (inclúyase hospitales y centros de salud públicos y del Instituto Ecuatoriano de Seguridad Social, IESS) por los servicios prestados a sus afiliados, ya sean de tipo hospitalario o ambulatorio. Por ejemplo, si una persona tiene un plan de salud privado y va a una cita médica en algún hospital del IESS y luego se hace una operación en el mismo hospital, todos esos gastos se descontarán de la cobertura anual de su seguro privado y la empresa de medicina prepagada deberá cancelar al Estado los valores producto de estos servicios hasta el monto anual que tuviese contratado el afiliado. Hubo empresas que incluso decidieron dar por terminados contratos corporativos para evitar el pago de estos valores.

4.1.2 SITUACIÓN ECONÓMICA DE LAS EMPRESAS DE MEDICINA PREPAGADA

A continuación, se verán datos del 2016 y 2017 (los años en los que se implantó la ley y donde hubo muchísimos debates al respecto) lo que mostrará un poco el efecto inmediato que tuvo la ley sobre las compañías de medicina prepagada. Para el año 2017, el sector representaba 497,8 millones de dólares según datos de Superintendencia de Compañías, Valores y Seguros, de estos, el 80 % se concentra en las 5 principales compañías. Las 5 compañías más grandes de medicina prepagada presentaron un crecimiento anual de 9,9 % [41].

La tabla 4.1 muestra como las principales empresas crecieron durante el 2016 y 2017.

La empresa líder en el mercado en estos años sufrió un decremento autoinfligido de 14,6 % producto de que la ley entraría en vigencia y antes de que esto suceda la empresa decidió voluntariamente dar por terminados algunos contratos corporativos que representaban un peligro cierto para la sostenibilidad de la empresa ya que entre otras cosas la ley exigía el pago al Estado por las atenciones recibidas por sus afiliados en centros de salud públicos, el mismo que se estimó en cerca de 70 millones anuales para todo el sector, en primera

Tabla 4.1: Principales Empresas de Medicina Prepagada 2016 a 2017
Fuente: Revista Ekos, 2018

EMPRESA	2016		2017		Crecimiento de Ingresos	Rentabilidad (ROI)	Participación
	Ingresos	Utilidad	Ingresos	Utilidad			
SALUDSA SISITEMA DE MEDICINA PRE-PAGADA DEL ECUADOR S.A.	150.174,28	12.651,09	128.319,79	16.697,84	-14,6 %	13,0 %	25,7 %
BMI IGUALAS MÉDICAS DEL ECUADOR S.A.	87.409,67	439,93	99.501,10	5.782,07	13,8 %	5,8 %	19,9 %
MEDICINA PARA EL ECUADOR MEDIECUADOR-HUMANA S.A.	39.300,03	23,72	62.080,25	3.616,98	58,0 %	5,8 %	12,4 %
*ECUASANITAS S.A.**	56.286,26	2.840,50	61.273,78	7.390,79	8,9 %	12,1 %	12,3 %
BEST DOCTORS S.A. EMPRESA DE MEDICINA PREPAGADA	30.896,82	- 663,75	48.763,76	4.098,00	57,8 %	8,4 %	9,8 %

**Datos Entregados por la institución*

Datos en miles de dólares

instancia.

Esta situación fue aprovechada por Humana S.A. principalmente (y otras compañías en menor medida) para absorber estos contratos, negociarlos adecuadamente y conforme a la ley presentar tarifas que les permitan dar servicios a estos clientes y mantener una utilidad para la empresa, producto de esta hábil maniobra creció en un 58 %.

El sector de medicina prepagada, como ya se dijo, creció un 9,9 % contrario a las primeras prospecciones que esperaban un incremento mucho menor debido al impacto de la ley.

Uno de los requisitos para operar era cumplir con un mínimo de capital suscrito de USD 1 millón, el mismo que debía cumplirse máximo hasta octubre del 2017. Hasta antes de ese momento 25 firmas estaban en capacidad de ofrecer planes de medicina prepagada, pero a partir de la fecha señalada, solo constaban 18 empresas registradas y autorizadas para brindar el servicio, según una circular de la Superintendencia de Compañías, Valores y Seguros (Oficio No. SCVS-INS-DNA-2018-00008603-OC, Superintendencia de Compañías, Valores y Seguros). Seis de las siete empresas restantes entraron en procesos de disolución y liquidación, y una se mantenía en procesos de solucionar la advertida falta de capital pagado previo a su aprobación; Esa empresa no podía celebrar contratos o captar afiliados mientras no se hubo solventado el tema.

El mercado se contrajo por la misma razón y muchos de los clientes de estas empresas simplemente no recompraron un plan en otra empresa debido a los nuevos precios producto de las exigencias de la ley.

En el artículo 10, la ley exige que las empresas tengan solvencia, inversiones obligatorias y montos de reservas por servicios prestados y no reportados y servicios prestados y reportados. Únicamente las reservas corresponden al 10 % de las primas anuales y considerando que la rentabilidad máxima de las empresas principales fue de 13 %, esto significaba que la consolidación de estas reservas se tomaría en el mejor de los casos toda la utilidad recibida por una empresa en un año. El plazo para la constitución de reservas fue de 3 años (2017 a 2019) en primera instancia, pero luego se estiró hasta el 2021. Algunas empresas ya han reportado importantes bajas en su utilidad (perdidas inclusive) debido a la conformación de dichas reservas.

La ley también exige algunos beneficios extras para quienes tengan el servicio de medicina prepagada y esto por supuesto supone un incremento en la siniestralidad debido a que muchos de estos beneficios son cargados directamente a la tarifa y no se paga copago, lo que a ojos de los afiliados se vería como gratuito, aunque no sea así. Considerando la poca difusión que estos beneficios han tenido y el hecho de que los usuarios tardan en entender la manera adecuada de exigir estos derechos era obvio que, en los primeros años de la ley, pocas personas hicieran uso de ellos, pero esto cambiará y cada vez más clientes harán uso de los mismos. La consecuencia de todo lo anterior es que paulatinamente la siniestralidad sube y los verdaderos efectos de todo lo promulgado en la ley recién se comenzaron a vislumbrar en el 2018 y se esperaría que para el 2020 todos quienes tengan un plan de medicina prepagada sepan a que atenerse y que exigir. Por lo anterior, solo a partir del año 2019 las empresas de medicina prepaga empezaron a tener plena conciencia de como fluiría el sector en adelante.

4.1.3 QUE SE ESPERA Y PORQUÉ ESTÁN IMPORTANTE EL PRESENTE ESTUDIO

El 15 noviembre del 2019, la Superintendencia de Compañías, Valores y Seguros informó mediante el oficio No. SCVS-INS-2019-00085438-OC [1] que diecinueve (19) compañías estaban registradas y autorizadas para emitir contratos de servicios de atención integral de

[1] link: https://portal.supercias.gob.ec/wps/wcm/connect/663487aa-7489-4acb-8ed3-58868f05d81a/SCVS-INS-2019-00085438-OC+PUBLICO+EN+GENERAL+-+SE+INFORMA+EL+LISTADO+DE+LAS+EMPRESAS+AUTORIZADAS+A+OFRECER+COBERTURAS+DE+SEGUROS+EN+ASISTENCIA+MEDICA+..pdf?MOD=AJPERES&CACHEID=663487aa-7489-4acb-8ed3-58868f05d81a

salud prepagada.

El 2019 se avizoró como un año poco prometedor para la economía, esto se pudo ver en el sector de la salud principalmente en las estimaciones conservadoras, en relación a inversiones, proyección de ventas y control del gasto, que hicieron las cabezas principales de hospitales y clínicas del país. Además, todos los presupuestos y proyecciones fueron modificados para enfrentar escenarios más conservadores, con inversiones de salud que se esperaban sean bajas y, algunos casos, nulas.

En general, en ese año se esperaba una disminución del gasto corriente traducido en reducción de personal en el sector público principalmente y que a su vez implicaba que habría una reducción del circulante y de las líneas de crédito y por parte de la banca se esperaba un incremento de la tasa de captación. Ecuador estaba a la espera de la respuesta por parte de FMI (Fondo Monetario Internacional) respecto al refinanciamiento de la deuda, con lo que tendríamos un respiro necesario en ese entonces y también hoy.

La contracción de la economía hace que los usuarios de servicios médicos vuelvan al control del costo beneficio ya que, al haber menos efectivo en circulación, menos créditos y menos empleo, sobre todo, las personas empiezan a evaluar la necesidad del gasto. Como ejemplo podemos decir que las cirugías no emergentes podrían ser postergadas, las consultas médicas solo se harían en los casos estrictamente necesarios, la mayoría de veces con médicos internos de las empresas, y ni hablar de reconsultas. Además, los usuarios tienden a preferir los procedimientos que puedan resultar más baratos (Siempre y cuando se tenga una garantía de seguridad y calidad), por ejemplo ciertos tratamientos para el cáncer, que se realizan usualmente de manera hospitalaria, pueden realizarse de manera ambulatoria a menor costo, pero con ciertas incomodidades.

El panorama para las empresas de medicina prepagada no fue diferente. Además, para el año 2020 se espera un incremento en la siniestralidad, que es el principal indicador actuarial. Puede parecer contradictorio que los hospitales y clínicas disminuyan sus ingresos y que las empresas de medicina prepagada esperen un incremento de siniestralidad porque ¿Cómo es posible que este tipo de empresas gaste más, si los hospitales dicen que van a tener menos ingresos?

Lo anterior, se sustenta en el hecho de que, al haber menos circulante, menos créditos, menos empleo y en general menos ingresos, entonces:

1. Las personas no podrán permitirse tener un plan de medicina prepagada o en su defecto tratarán de que este plan se ajuste a sus ingresos.

2. Los afiliados en general intentarán sacar el máximo provecho a su plan, y por su puesto esta no es una situación normal.

3. Por obvias razones los clientes que no necesiten el servicio o que en su opinión no es imprescindible serán los que opten por desertar, mientras que los clientes que lo requieren, y que por lo general son los que más gastos médicos presentan, son los que, entre comillas, se verán obligados a mantener el servicio.

Es decir, por un lado las primas que reciben las empresas se reducirán y por otro los gastos de algunos afiliados se incrementaran y además de eso los afiliados que desertarán serán los considerados mejores clientes para la empresa.

Por supuesto, esto hará que se deban mejorar los procesos, ya sea con mejor personal o capacitando al que se tiene para tener un control más adecuado del gasto y además mejorar los convenios con los proveedores de modo que se tenga mejores precios o mejores prestaciones.

Captar nuevos clientes en este tipo de escenarios será muy difícil, las ventas serán extremadamente disputadas y debido a que la ley tiene muchas prohibiciones, competir mejorando o dando un plus a nuestros productos se volverá muy complicado, de ahí que retener a los que ya son nuestros clientes será muy importante.

En todo lo anterior radica el valor de implementar el presente modelo; prevenir que los afiliados por alguna razón quieran dar por terminada la relación contractual con la empresa es necesario, pero para esto es imperativo identificar a estos afiliados que están a punto de desertar o lo están pensando hacer en el corto plazo. Mientras más eficiente sea la predicción mucho mejor serán los perfiles y las estrategias podrán ser aplicadas de manera más particular.

4.2 EXPERIENCIAS Y BENEFICIOS DEL ANÁLISIS DE SUPERVIVENCIA EN EL TRATAMIENTO DE LA DESERCIÓN DE CLIENTES DESDE LA PERSPECTIVA DE DIFERENTES INDUSTRIAS

En este apartado se expondrá dos trabajos en los que se aplicó el análisis de supervivencia con la regresión de Cox para predecir la deserción de clientes, identificar posibles causas y relaciones entre variables.

El primer artículo que se expondrá es “Applying Survival Analysis to Telecom Churn Data” por Masarifoglu y Buyuklu [42]; la metodología usada en la publicación antes mencionada, además de la usada en otras fuentes como la investigación de Balboa [43], influyó significativamente en la metodología usada en el presente trabajo. El segundo trabajo que se expondrá es el del Análisis de deserción de clientes en una empresa de Servicios Financieros [44] el cual es importante ya que no solo valida el uso del modelo en cuestión para atacar el problema de la deserción de clientes, sino que también nos muestra algunas de las dificultades que tuvieron los autores.

Ambos artículos darán una idea de lo que podemos esperar al usar el modelo de riesgos proporcionales, los beneficios, las dificultades y los puntos clave de cada una de las etapas construcción y análisis de resultados.

4.2.1 APLICACIÓN DEL ANÁLISIS DE SUPERVIVENCIA A DATOS DE DESERCIÓN EN TELECOMUNICACIONES

En este apartado se analizará en trabajo de Masarifoglu y Buyuklu como lo habíamos mencionado antes. Como sugiere el título, se analiza la deserción de clientes en una empresa de telecomunicaciones de la cual se extrae la data y lo que se requería saber es cómo influyen en la deserción de clientes variables como incentivos, campañas, tarifas, la edad, método de pago, entre otras.

Los autores tomaron una base de 10365 cliente seleccionados aleatoriamente y a los cuales se les dio seguimiento por un año desde el primero de enero de 2015 hasta el 31 de diciembre del mismo año, periodo durante el cual 2654 desertaron (Estos representaron el 26,5 % de la población inicial). Las covariables seleccionadas fueron: Campañas (tomada

como una variable binaria o 'dummy'), la tarifa (que es una variable ordinal que puede tomar 4 valores), la permanencia (que es el tiempo que el cliente a estado con la empresa, medida en meses), la edad (la cual fue transformada en una variable ordinal que puede tomar 4 valores posibles) y el autopago (Que quiere decir si el cliente pagaba automáticamente a través de débito bancario). La variable campañas fue eliminada por contradecir la hipótesis de riesgos proporcionales, mientras que las demás variables entraron al modelo.

La permanencia y la edad fueron las dos covariables que más influyen en la decisión de abandonar la empresa. Si se mantienen constantes todas las variables y se incrementa un mes la permanencia de un cliente entonces el riesgo de deserción se reduce en un 5 % . Para la edad, manteniendo constantes todas las demás variables, un cliente en la categoría 2 tiene un 11 % menos de probabilidades de abandonar que un cliente en la categoría 1, en cualquier momento durante un año.

Claramente podemos ver que los resultados son más que satisfactorios y que se logró no solo identificar las principales relaciones entra las variables y la probabilidad de deserción, sino que también se puede predecir de manera bastante acertada la deserción de los clientes. Esto les da a los departamentos de marketing una herramienta valiosa para priorizar clientes y desarrollar herramienta personalizadas de retención de clientes.

4.2.2 ANÁLISIS DE DESERCIÓN DE CLIENTES EN UNA EMPRESA DE SERVICIOS FINANCIEROS

En esta sección se propone revisar como el modelo de riesgos proporcionales de Cox fue una buena elección al momento de evaluar la deserción de clientes en una importante empresa de servicios financieros en Europa.

Los autores, Van den Poel y Lariviere, documentaron cada uno de sus hallazgos y además de esto las dificultades y limitaciones que tuvieron al momento de realizar el estudio, lo que vuelve a este trabajo doblemente importante.

El análisis presentado en este artículo se basa principalmente en la influencia sobre la deserción del cliente de las covariables agrupadas en cuatro grandes grupos: comportamiento del cliente, percepciones del cliente, demografía del cliente y entorno macro. Aquí aparece la primera dificultad, ya que las variables de percepción del cliente se recolectaron a través de encuestas y otras como las demográficas se obtuvieron de la base de datos de la em-

presa patrocinadora. Juntar estas variables fue complicado y requirió alto nivel de detalle, según los autores.

Usando el modelo de riesgos proporcionales pudieron determinar que las características demográficas y los cambios ambientales tienen mucha relación con la deserción y un gran impacto en la retención. Los predictores de comportamiento tienen un impacto limitado en la decisión del cliente de terminar su relación contractual con la empresa a través del número total de productos que posee y el tiempo de compra, sino que más bien los predictores de comportamiento “cuántos” y “qué tan recientemente” reclaman la atención. La propiedad específica del producto, la propiedad de la tarjeta, el hecho de realizar operaciones bancarias en el hogar o por teléfono no afectan las tasas de retención. En resumen, la industria de servicios financieros puede reducir las tasas de abandono al ofrecer a los clientes “nuevos” incentivos para quedarse y no tanto nuevas facilidades.

Dicho esto, cabe recalcar que, en palabras de los propios autores, “La técnica es apropiada para usar debido a la naturaleza dependiente del tiempo de la mayoría de las covariables y porque se cumple el supuesto básico de proporcionalidad”.

Los autores también mencionan que sería interesante considerar a los clientes ’dormidos’, que solo tienen una pequeña cantidad en una cuenta, así como los clientes abandonados o enfocarse en el evento de clientes que tienen defectos ’parciales’ en lugar de una deserción ’total’. En particular, esta es una de las conclusiones a las que este trabajo también llegará, como se podrá ver en las secciones finales, y es que especificar, diversificar y mejorar la definición de deserción parece ser una necesidad absoluta y que en este momento no puede atenderse debido a la falta de datos o la falta de certeza de los mismos.

4.3 DESCRIPCIÓN DE LOS RIESGOS DE LA CAÍDA DE CARTERA EN EMPRESAS DE MEDICINA PREPAGADA

La caída de cartera en las empresas de medicina prepagada, debida principalmente a la deserción de los clientes, afecta esencialmente de tres maneras diferentes:

1. Perdida del ingreso
2. La mala propaganda
3. Perdida de la capacidad de reacción ante casos costosos

Cada uno de estos puntos será analizado por separado.

4.3.1 PERDIDA DEL INGRESO

Por supuesto, la afectación que salta a la vista es la pérdida del ingreso producto del pago de las cuotas del plan adquirido por el cliente. Se podría pensar, a primera vista, que es la única que importa o la más importante, pero esto no es así debido a la naturaleza de la industria.

Como veremos más adelante, hay dos daños colaterales en particular para las empresas de medicina prepagada y aunque la pérdida del ingreso siempre es el principal aspecto negativo que se quiere evitar con el abandono del cliente, no siempre es el que más afecta a la empresa.

4.3.2 LA MALA PROPAGANDA

No en todos los casos la pérdida del ingreso es el principal perjuicio para la empresa, por ejemplo, tomemos el caso de Dave Carrol. Carrol es un músico canadiense que, durante 9 meses, exigió una indemnización por parte de la aerolínea United Airlines debido a que en un viaje que él hizo desde Halifax a Nebraska, en un vuelo de la mencionada aerolínea, el personal de embarque lanzó su estuche de guitarra por los aires y al llegar a su destino, Dave se encontró con su guitarra rota. Cansado de ser ignorado y luego de recibir una negativa basada en las políticas de la empresa, Dave se desquitó publicando una canción de protesta en julio del 2009, la misma que desde luego desprestigiaba a la aerolínea. El video tiene a la fecha más 19 millones de visualizaciones en YouTube y en tan solo 4 días desde su publicación, las acciones de la aerolínea bajaron un 10 % causando una pérdida de 180 millones de dólares. Perder un cliente puede significar mucho más que eso.

Ninguna empresa está exenta de perder clientes, sin embargo, para las empresas de medicina prepagada el riesgo es mucho mayor ya que si el cliente decide terminar la relación contractual, debido a alguna mala experiencia relacionada con la empresa o con los prestadores de servicios médicos, los clientes, por lo general, acaban cancelando todos los contratos de todos los productos que tengan y muy comúnmente comentan su situación con amigos, familiares y conocidos. Para la industria de la medicina prepagada y para muchas otras, la principal publicidad es la que hacen sus propios clientes, y si los clientes hablan

mal de la empresa es poco probable que las personas a su alrededor estén dispuestos a contratar sus productos.

Un estudio realizado en Reino Unido por NewVoiceMedia, reveló que un 56 % de los encuestados que ha pasado por una experiencia negativa con una marca o compañía, jamás volvería a comprar nada relacionado con esta. De igual manera, un 27 % dijo que estaba dispuesto a contarle a sus amigos y conocidos y un 19 % lo divulgaría por redes sociales [45].

4.3.3 PERDIDA DE LA CAPACIDAD DE REACCIÓN ANTE CASOS COSTOSOS

Antes de empezar a disertar sobre este tema, me gustaría recordar lo que dice la ley fuerte de los grandes números. La ley fuerte de los grandes dice que:

Teorema. Sea $(X_i, i \geq 1)$ una sucesión de variables aleatorias independientes idénticamente distribuidas y sea $E[X_1] = \mu$, entonces

$$\frac{X_1 + X_2 + ... + X_n}{n} \to \mu \quad cuando\, n \to \infty$$

con probabilidad 1 [46].

En palabras simples, la ley fuerte de los grandes números afirma que cuando la cantidad de observaciones de una muestra aumenta, la media de la muestra se acerca al valor esperado.

En la industria de seguros y de medicina prepagada podemos decir que a medida que aumenta el número de unidades de exposición (asegurados, expuestos o afiliados), la probabilidad de que la pérdida real por unidad de exposición sea igual a la pérdida esperada por unidad de exposición es mayor. Para decirlo en términos económicos, hay rendimientos a escala en la producción de seguros.

En términos prácticos, esto significa que es más fácil establecer la prima correcta, y por lo tanto reducir la exposición al riesgo, a medida que se emiten más pólizas dentro de una clase de seguro o plan de medicina prepagada determinado. Es mejor que una compañía de seguros emita 1000 pólizas de seguro contra incendios en lugar de 200, suponiendo una distribución de probabilidad estable e independiente para la exposición a pérdidas.

Para verlo de otra manera, suponga que una compañía de medicina prepagada descubre que cinco de cada 100 personas sufrirán una lesión grave y costosa durante un año determinado. Si la compañía asegura solo a 20 personas, se enfrenta a riesgos mucho mayores que si pudiera asegurar a 1000 personas, ya que en 20 personas se espera que solo una sufra una lesión costosa, pero ese número puede cambiar fácilmente, por ejemplo, podría ser dos personas y de esta manera el costo se duplica con una mínima desviación. La compañía puede estar más segura de que 1000 asegurados pagarán colectivamente primas suficientes para cubrir las reclamaciones de cincuenta clientes que sufren lesiones graves y aunque sufran cinco casos más, el desvío del costo no es tan grave como en el primer caso.

Por todo lo anterior es fácil ver que la pérdida de clientes incrementa el riesgo de las empresas de medicina prepagada y que es necesario tener la mayor cantidad de afiliados posible. Perder muchos clientes es un riesgo y bajo circunstancias extremas puede significar la banca rota de la compañía.

Obviamente estos riesgos son atenuados con la contratación de reaseguros, sin embargo, esto también eleva la prima que debe pagar el afiliado y vuelve menos competitivo al producto, sin mencionar que hay planes de medicina prepagada para los cuales es impensable contratar un reaseguro debido a que su bajo costo hace que un mínimo incremento los saque definitivamente del mercado.

4.4 ANTECEDENTES DEL CONTROL DEL RIESGO DE DESERCIÓN EN LAS EMPRESAS MEDICINA PREPAGADA EN AMÉRICA LATINA Y EL MUNDO

Sobre el control de la deserción en las empresas de medicina prepagada existe muchísima literatura. Podemos encontrar desde trabajos como el de Jefferson Camelo Soares (en colaboración con otros 6 autores) “How to Avoid Customer Churn in Health Insurance/Plans? A Machine Learn Approach” [47] cuyo objetivo principal es desarrollar un enfoque para predecir la cancelación opcional del contrato en un plan o seguro de salud privado y ayudar a las compañías a prevenir esas cancelaciones, hasta proyectos completos de fidelización de clientes que carecen de todo análisis de datos y que se basan en el puro conocimiento técnico y empírico, dejando completamente de lado su propia realidad para asumir que lo

que funcionó en otra población de clientes funcionará para la suya. Otros cuantos trabajos se enfocan netamente en estrategias de retención, es decir, son metodologías reactivas, cuya falta de precisión trae consecuencias económicas indeseables cuando los gastos de retención superan lo presupuestado y no tienen los resultados que se aspiraban.

Los cierto es que tanto las estrategias reactivas como los planes de fidelización despersonalizados son cosa del pasado. En la era del big data el uso de metodologías obsoletas y la falta de modernización son errores que se pagan caro.

Ya en el 2009 se podían encontrar trabajos como el de Jin Su, Kimberly Cooper, Tina Robinson y Brad Jordan de 'BlueCross BlueShield' [2] cuyo objetivo era el desarrollo de modelos de retención para ayudar a empresas de seguros de salud a identificar los impulsores clave para aumentar la retención y enfocar la comunicación a los clientes "correctos". Este documento abordó las dificultades que enfrenta la industria de la salud en la construcción de un mercado de datos de series temporales y desarrolló un modelo predictivo de retención (Mediante una regresión logística) que podía entregar puntajes de retención para cada cliente. Lograron identificar los principales factores que afectan la retención a través del proceso de modelado y también se discute en el documento algunas de sus aplicaciones [48].

En Europa, tenemos investigaciones como la realizada por la compañía holandesa de seguros de salud CZ [49]. En este trabajo, la predicción de la deserción de clientes a partir de variables objetivas en CZ se investiga sistemáticamente utilizando técnicas de minería de datos. Para identificar importantes variables y características de deserción, se entrevistó a expertos dentro de la empresa y además de esto se analizó y examinó la literatura existente al respecto. En este punto quisiera hacer un paréntesis, ya que, por experiencia personal, si bien los expertos son una excelente fuente de conocimiento, usualmente son los colaboradores que hablan y reciben las quejas y sugerencias directamente del cliente quienes son la mejor fuente de enriquecimiento de las investigaciones de este tipo. Además, se identificaron cuatro técnicas prometedoras de minería de datos para el modelado de predicciones que son la regresión logística, el árbol de decisión, las redes neuronales y los modelos SVM. Después de la evaluación del rendimiento, la regresión logística con un conjunto de entrenamiento 50:50 (no desertor: desertor) y redes neuronales con una dis-

[2] BlueCross BlueShield es una asociación de 36 empresas de seguros médicos, actualmente con sedes en Chicago y Washington. Según un artículo de *thebalance.com* tiene la mejor y más grande red de prestadores. Otro artículo de *insure.com* colocó a 4 de sus compañías en el top 10 de compañías de seguros médicos en los Estados Unidos.

tribución 70:30 (no desertor: desertor) se desempeñaron mejor. En el caso ideal, se puede identificar al 50 % de los desertores contactando solamente al 20 % de la población, con estos resultados el análisis de costo-beneficio arrojó que si hay un equilibrio entre los costos de contactar a estos clientes y los beneficios de la retención de clientes.

Si bien la búsqueda de trabajos relacionados con la deserción, retención o fidelización de clientes en empresas de seguros de salud o medicina prepagada a nivel mundial fue un poco complicada, en América latina fue un verdadero desafío. Pocas investigaciones se han hecho al respecto y la mayoría son planes o proyectos de retención o fidelización de clientes basados básicamente en encuestas. En resumen, se puede decir que las empresas de medicina prepagada en los países latinoamericanos en general sufren un retraso considerable respecto a este tema, probablemente por la falta conciencia de las consecuencias que resultan de no dar un seguimiento adecuado a nuestros clientes y el desconocimiento de los enormes beneficios de hacerlo.

4.5 MODELO DE RIESGOS PROPORCIONALES PARA LA PREDICCIÓN DE LA DESERCIÓN DE CLIENTES EN UNA EMPRESA DE MEDICINA PREPAGADA EN ECUADOR

Explicaremos pasa a paso el desarrollo de un modelo de riesgos proporcionales para la predicción de la deserción de clientes, es decir, la no continuidad del cliente como afiliado a un plan de medicina prepagada de la empresa Humana S.A., patrocinadora del presente trabajo. Previo a esto debe recalcarse que algunas pequeñas partes de la información y del modelo final serán intencionalmente ocultadas para evitar la divulgación de información sensible o confidencial.

La empresa Humana S.A. ha compartido los datos de los clientes de la línea de negocio individual (excluyendo datos personales) entre los que podemos hallar variables demográficas, geográficas, comportamiento de pago [3], uso del servicio, preferencias de prestadores, enfermedades padecidas, movimientos, frecuencia, siniestralidad y por supuesto información sobre la desafiliación.

[3] Se refiere al comportamiento de pago exclusivamente con la empresa patrocinadora y no en el sistema financiero en general.

4.5.1 DESCRIPCIÓN DE LOS SUJETOS DE ESTUDIO

Previamente se mencionó que para el presente estudio se seleccionó los contratos de la línea de negocio individual, por lo que describiremos cómo se tomo la información y de qué manera se unió las bases brutas.

Los clientes de la línea de negocio individual suelen ser personas solas (mayores de edad) o familias, para este estudio debemos aclarar que cada sujeto de estudio se refiere a un contrato de esta línea de negocio y que no se refiere a un afiliado o persona, es decir, los datos fueron tomados por contrato y no por persona y dentro de cada contrato pueden estar incluidos una o más personas. La razón de esta decisión es que normalmente la desafiliación se da para todo el contrato y no es solo un afiliado el que decide de manera personal salir del contrato [4]. Cada contrato tiene un solo plan de medicina prepagada, con raras excepciones que fueron desechadas del estudio, y cada afiliado tiene una prima diferenciada por edad (rango etario) y sexo, tal como lo solicita la ley.

Existen casos en los en un contrato solo constan menores de edad como beneficiarios y están representados por un titular con mayoría de edad pero que no quiere tener acceso a los servicios por propia voluntad y por ende tampoco paga una prima (Esto es muy común cuando el representante tiene un plan de medicina prepaga corporativo, es decir, otorgado por su empleador y no considera necesario otro plan). A esta figura se le llama titular sin beneficios. Esta misma figura se usa cuando hay personas con mayoría de edad pero que no puede representarse solas, como es el caso de personas con discapacidad, y cuyos representantes no desean ser incluidos en el contrato.

La base de datos que se entregó para este estudio estaba desglosada por afiliado y no por contrato por lo que hubo que ordenar la información y crear nuevas variables (106 variables fueron creadas).

4.5.2 DEFINICIÓN DEL EVENTO DE INTERÉS

El evento de interés de este estudio es la desafiliación del contrato en su totalidad lo que en principio parece bastante simple pero no lo es del todo y lo se lo explicará a continuación.

[4] Esto se explicará de mejor manera en el siguiente apartado.

La ley de medicina prepagada requiere que para que una persona acceda a los servicios de una empresa de medicina prepagada, valga la redundancia, debe firmarse un contrato genérico, el mismo que es validado por los entes de control, y el futuro afiliado también debe entregar una declaración de salud [5]. Este contrato debe tener un máximo de duración de un año y luego de este se puede renovar automáticamente por un año más a menos que el afiliado decida no continuar con el servicio.

Para desafiliarse, la ley requiere que el afiliado entregue una carta a la empresa donde exprese su deseo de dar por terminada la relación contractual. Esta carta debe presentarse con tiempo prudencial antes de la fecha de corte del pago de la siguiente cuota mensual (cinco días al menos) y no es necesario que el contrato tenga una duración total de un año, sino que el afiliado puede presentar la carta en cualquier momento.

La empresa patrocinadora da un extensivo seguimiento al proceso mediante una llamada al titular del contrato en la que se trata de rescatar al cliente mediante diferentes estrategias reactivas y una encuesta post deserción [6] cuando esto no se logra.

Todo el proceso es registrado en una base de datos donde se tiene tres fechas:

- La **Fecha de aprobación** que es cuando el ejecutivo de servicio al cliente ya realizó la llamada y no pudo persuadir al cliente, por lo tanto determina la ejecución de la desafiliación.
- La **Fecha de proceso** que es cuando se registra la desafiliación en el sistema.
- La **Fecha de movimiento** que es fecha hasta la que legalmente llegó el contrato.

En ocasiones el proceso ha tenido que hacerse de manera retroactiva y esto afecta a las dos primeras fechas, pero no a la tercera. Es por esto que se considera la tercera fecha como fecha de deserción del contrato.

Sería deseable tener la fecha en la que el afiliado entrego la carta de desafiliación, pero desafortunadamente esta fecha no se registra.

En base a todo lo anterior, el evento de interés se define como **Desafiliación**, entendiéndose como el instante en el que legalmente se considera finalizado el contrato por decisión

[5] Este es un documento donde el afiliado declara sus enfermedades preexistentes

[6] Esta encuesta está siendo aplicada desde hace relativamente poco tiempo por lo cual los datos de la misma no fueron incluidos en este estudio.

expresa del cliente.

4.5.3 VENTANA DE MUESTREO

Los datos para la predicción fueron tomados en un periodo de 4 años desde el 1 de enero de 2016 hasta el 31 de diciembre de 2019.

Los tres primeros años se usaron para obtener la información de las covariables y crearlas y el último año se usó para verificar el estado del sujeto durante ese tiempo, es decir, si le ocurrió el evento de interés o no.

Se cortó la información de los contratos al 1 de enero de 2019, es decir, se tomó como sujetos de estudio a todos los contratos vigentes a esa fecha. Se tomo toda la información de los contratos desde el 1 de enero de 2016 de manera de crear las variables predictivas y luego se dio seguimiento de 1 año para verificar la ocurrencia o no ocurrencia del evento de interés. En caso de no ocurrir se censuró los datos y en caso de ocurrir se tomó el tiempo de permanencia desde el 1 de enero de 2019 hasta la fecha de desafiliación del contrato.

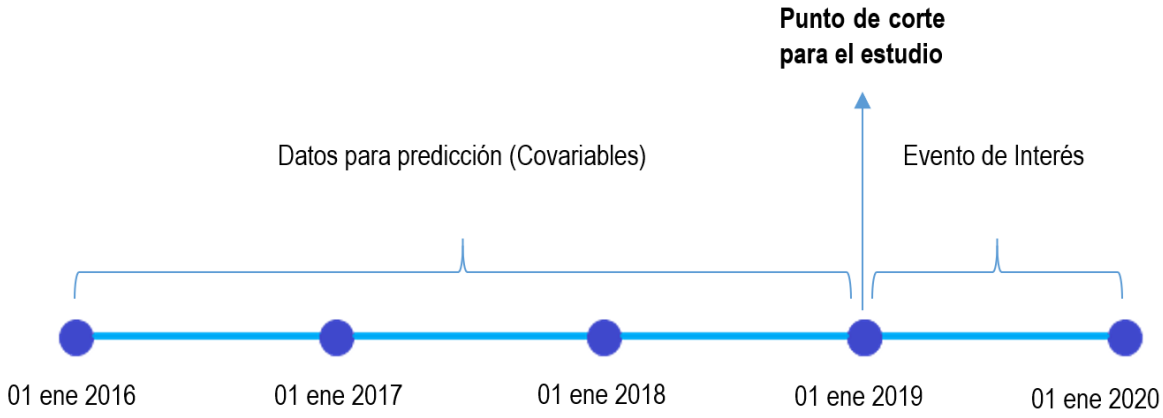


Figura 4.1: Diagrama de diseño de obtención de datos

4.5.4 ANÁLISIS DESCRIPTIVO DE BASE BRUTA

La base bruta recibida por Humana S.A. fue transformada en una base por contrato con 106 variables creadas; todos los datos fueron transformados a variables numéricas y a continuación se muestra un resumen descriptivo de todas las variables.

En base a la tabla 4.2, podemos ver que hay muchísimas variables que no tienen suficientes

Tabla 4.2: Descripción de Variables de la base bruta

Variable	Mínimo	Primer Cuartil	Mediana	Media	Tercer Cuartil	Máximo	Valores Perdidos
CONTRATO							
CENSURA							
TIEMPO	0	249	366	296,4	366	366	0
PRESENTADO	0	0	0	155,29	84,88	20394,31	0
PAGADO_NETO	0	0	0	109,32	49,26	14503,79	0
PORCENTAJE_COBERTURA	0	0,5793	0,6999	0,6559	0,7736	1,0878	3062
PRESENTADO_3M_PA	0	0	0	58,95	0	19870,88	0
PAGADO_NETO_3M_PA	0	0	0	32,32	0	13079,73	0
PORCENTAJE_COBERTURA_3M_PA	0	0,144	0,482	0,435	0,685	1	9945
PRESENTADO_6M_PA	0	0	0	149,6	50	76164,8	0
PAGADO_NETO_6M_PA	0	0	0	81,28	0	34412,87	0
PORCENTAJE_COBERTURA_6M_PA	0	0,198	0,469	0,428	0,65	1	8688
PRESENTADO_9M_PA	0	0	0	232	131,2	77028,2	0
PAGADO_NETO_9M_PA	0	0	0	125,9	32	34862,3	0
PORCENTAJE_COBERTURA_9M_PA	0	0,208	0,458	0,413	0,616	1	7924
PRESENTADO_12M_PA	0	0	0	313,2	205,4	77647,1	0
PAGADO_NETO_12M_PA	0	0	0	173,8	72	35336,4	0
PORCENTAJE_COBERTURA_12M_PA	0	0,244	0,477	0,43	0,626	1	7426
PRESENTADO_3M	0	0	0	155,29	84,88	20394,31	0
PAGADO_NETO_3M	0	0	0	109,32	49,26	14503,79	0
PORCENTAJE_COBERTURA_3M	0	0,57	0,7	0,64	0,776	1	6961
PRESENTADO_6M	0	0	53,57	430,09	319,99	76903,41	0
PAGADO_NETO_6M	0	0	28,36	306,59	203,54	51097,06	0
PORCENTAJE_COBERTURA_6M	0	0,577	0,703	0,65	0,782	1,088	4824
PRESENTADO_9M	0	0	121	685,3	551,2	77885,4	0
PAGADO_NETO_9M	0	0	66,51	491,58	354,85	51465,96	0
PORCENTAJE_COBERTURA_9M	0	0,573	0,704	0,65	0,784	1,088	4034
PRESENTADO_12M	0	0	177,8	925,3	773,6	81493,3	0
PAGADO_NETO_12M	0	0	104,3	670	503,6	70251,1	0
PORCENTAJE_COBERTURA_12M	0	0,583	0,704	0,656	0,784	1,088	3649
RECORTE_12M	0	0	0	29,22	10	2984,75	0
RECORTE_9M	0	0	0	22,45	5,55	2744,04	0
RECORTE_6M	0	0	0	15,04	1,3	2744,04	0
RECORTE_3M	0	0	0	5,751	0	1761,83	0
REEMBOLSOS_CERO_PAGO_3M	0	0	0	0,07637	0	5	0
REEMBOLSOS_CERO_PAGO_6M	0	0	0	0,183	0	7	0
REEMBOLSOS_CERO_PAGO_9M	0	0	0	0,299	0	11	0
REEMBOLSOS_CERO_PAGO_12M	0	0	0	0,3669	0	12	0
MAX_TIEMPO_OCURRENCIA_PAGO_12M	1	37	50	51,61	63	273	3649
MAX_TIEMPO_OCURRENCIA_PAGO_9M	1	36	47	48,81	61	232	4034
MAX_TIEMPO_OCURRENCIA_PAGO_6M	0	34	43	45,17	57	133	4824
MAX_TIEMPO_OCURRENCIA_PAGO_3M	0	30	36	35,88	42	83	6961
EDAD_TITULAR	18	31	38	41,21	50	106	0
GENERO_TITULAR	0	0	0	0,4701	1	1	0
NUMERO_AFILIADOS	1	1	1	1,703	2	10	0
NUMERO_AFILIADOS_MENORES	0	0	0	0,544	1	7	0
NUMERO_AFILIADOS_5AÑOS_O_MENOS	0	0	0	0,2201	0	3	0
NUMERO_AFILIADOS_2AÑOS_O_MENOS	0	0	0	0,1004	0	2	0
INDIVIDUAL_FAMILIAR	1	1	1	1,377	2	2	0
CONYUGE	0	0	0	0,1853	0	1	0
TIEMPO_AFILIACION	3	214	501	843,8	986	4018	0
PRODUCTO_MH	0	0	0	0,4439	1	1	0
TOPE_PLAN	15000	30000	50000	53333	80000	150000	0
TITULAR_SIN_BENEFICIOS	0	0	0	0,1618	0	1	0
PRESENTADO_CRONICO_12M	0	0	0	317,5	127,6	81493,3	0
PAGADO_NETO_CRONICO_12M	0	0	0	229,68	80,02	70251,1	0
PORCENTAJE_COBERTURA_CRONICO_12M	0	0,617	0,725	0,679	0,808	1	6684
PRESENTADO_CPC_12M	0	0	58,77	661,16	413,54	81493,29	0
PAGADO_NETO_CPC_12M	0	0	31,32	486,19	273,54	70251,1	0
PORCENTAJE_COBERTURA_CPC_12M	0	0,6	0,718	0,67	0,801	1,088	4758
INCLUSIONES_3M	0	0	0	0,01103	0	3	0
INCLUSIONES_6M	0	0	0	0,02273	0	4	0
INCLUSIONES_9M	0	0	0	0,03208	0	7	0
INCLUSIONES_12M	0	0	0	0,03831	0	7	0
EXCLUSIONES_3M	0	0	0	0,01069	0	6	0
EXCLUSIONES_6M	0	0	0	0,02139	0	6	0
EXCLUSIONES_9M	0	0	0	0,03738	0	9	0
EXCLUSIONES_12M	0	0	0	0,04648	0	12	0
CAMBIO_PLAN_3M	0	0	0	0,07493	0	1	0
CAMBIO_PLAN_6M	0	0	0	0,1137	0	1	0
CAMBIO_PLAN_9M	0	0	0	0,1983	0	1	0
CAMBIO_PLAN_12M	0	0	0	0,236	0	1	0
MOVIMIENTOS_3M	0	0	0	0,2204	0	26	0
MOVIMIENTOS_6M	0	0	0	0,37	0	36	0
MOVIMIENTOS_9M	0	0	0	0,6152	1	36	0
MOVIMIENTOS_12M	0	0	0	0,7594	1	36	0
MORA_PROM_12M	0	0	0	2,014	1,571	132,5	4
MORA_MAX_12M	0	0	0	7,701	6	252	4
PAGOS_EFECTIVO_12M	0	0	0	0,01214	0	1	4
PAGOS_DEBITO_BANCARIO_12M	0	0	1	0,6015	1	1	4
PAGOS_TARJETA_CREDITO_12M	0	0	0	0,3861	1	1	4
MORA_PROM_6M	0	0	0	2,19	1,333	142,5	48
MORA_MAX_6M	0	0	0	6,18	4	169	48
PAGOS_EFECTIVO_6M	0	0	0	0,01032	0	1	48
PAGOS_DEBITO_BANCARIO_6M	0	0	1	0,6021	1	1	48
PAGOS_TARJETA_CREDITO_6M	0	0	0	0,3873	1	1	48
MORA_PROM_3M	0	0	0	2,3116	0,6667	64	113
MORA_MAX_3M	0	0	0	4,472	2	77	113
PAGOS_EFECTIVO_3M	0	0	0	0,00844	0	1	113
PAGOS_DEBITO_BANCARIO_3M	0	0	1	0,6031	1	1	113
PAGOS_TARJETA_CREDITO_3M	0	0	0	0,3882	1	1	113
INCREMENTO_RELATIVO_2018	-0,451	0	0,069	0,171	0,179	4,118	4462
INCREMENTO_ABSOLUTO_2018	-76,73	0	4,462	17,625	19,478	722,349	4462
DIAS_FIN_CONTRATO	0	102	167	177,7	247	362	0
NUEVO	0	0	0	0,2112	0	1	0
USO	0	0	1	0,7315	1	1	0
USO_6M	0	0	1	0,581	1	1	0
USO_12M	0	0	1	0,6822	1	1	0
USO_HOSPITALARIO	0	0	0	0,008504	0	1	0
USO_AMB_EMR_HSD	0	0	0	0,2443	0	1	0
TIEMPO_SIN_USO	3	53	92	166,8	185	1065	0
SINIESTRALIDAD_12M	0	0	0,06201	0,28204	0,2388	106,85233	0
SINIESTRALIDAD_6M	0	0	0,01525	0,15397	0,10001	106,08643	0
CANAL	0	1	1	0,8726	1	1	0
PRIMA_PE	0,08	45,65	64,41	79,28	97,29	387,53	0
INCREMENTO_ABS	-4493,419	0	0	9,836	7,86	4259,34	0
INCREMENTO_REL	-0,69223	0	0	0,10331	0,09999	28,70162	0

datos y probablemente deban ser eliminadas, sin embargo, antes de hacer esto es necesario considerar un problema que a priori no se había visto. Se sospecha que la deserción de clientes tiene un comportamiento diferente para clientes con un tiempo de afiliación menor a un año y los que tienen más de un año. Si se comprueba que esta sospecha es verdad, sería necesario crear dos modelos, uno para cada población.

Además, si se logra comprobar que existen diferentes comportamientos en estas dos poblaciones, entonces tendríamos que las variables que son significativas en un caso no lo son para el otro por lo que antes de eliminar variables y depurar es necesario verificar esta hipótesis.

4.5.5 DECISIÓN DE PARTICIÓN EN DOS MODELOS

Como ya se mencionó en el apartado anterior, se sospecha que los clientes con un tiempo de afiliación de un año o menos tienen un comportamiento diferente, en lo que se refiere a la deserción, a aquellos clientes que tienen más de un año de afiliación a la empresa. Para verificar la hipótesis vamos usar la prueba log-rank, que compara las funciones de riesgo de dos o más poblaciones.

Para la prueba log-rank se seleccionó una función de peso $W(t) = 1$ que según Klein y Moeschberger [6] es la mejor opción para verificar la hipótesis alternativa. La prueba se aplicará a través del siguiente código de lenguaje R, donde se usa una aproximación para el cálculo del p-valor debido a que el cálculo exacto es computacionalmente difícil. El código y su salida se muestran en el código R 4.1.

Código de R 4.1: Prueba log-rank para dos poblaciones diferenciadas por su tiempo de afiliación (1 año o menos y más de un año)

```
> library(coin)
> logrank_test(Surv(TIEMPO,CENSURA)~TIEMPO_AFILIACION, data=Base 20190101,
    distribution = approximate(B = 10000))

        Approximative Two-Sample Logrank Test

data:  Surv(TIEMPO, CENSURA) by TIEMPO_AFILIACION (<= 1 año, > 1 año)
Z = -15.865, p-value < 1e-04
alternative hypothesis: true theta is not equal to 1
```

Como el p-valor es menor que 0,05 entonces rechazamos la hipótesis nula, es decir, no

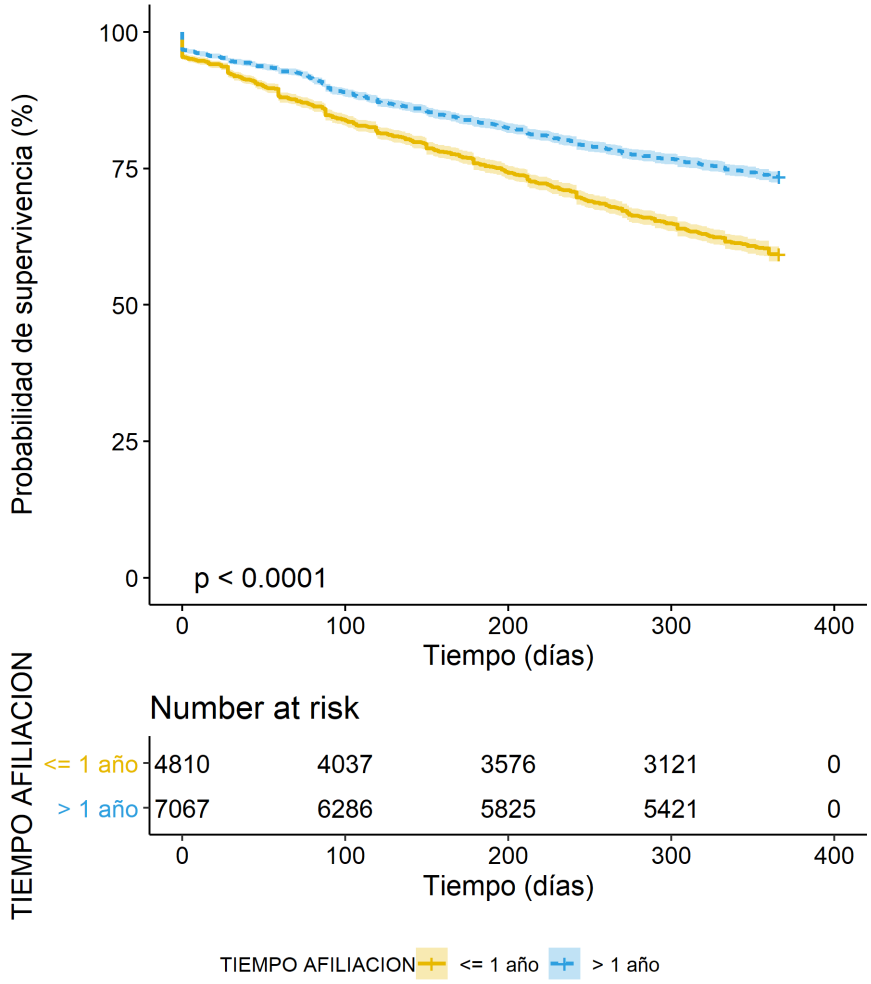


Figura 4.2: Comparativo de funciones de supervivencia de las poblaciones de cliente que tienen un tiempo de afiliación de más de un año (>1 año) y clientes con un tiempo de afiliación de un año o menos (<= 1 año)

existen pruebas que nos digan que las funciones de riesgo de las dos poblaciones son iguales.

La figura 4.2 nos muestra como las funciones de supervivencia de cada población son visiblemente diferentes, y es que los contratos con un año o menos de afiliación presentan menor probabilidad de supervivencia en el mismo tiempo que los contratos que tienen más de un año.

En conclusión, es necesario dividir las poblaciones con el objetivo de ser más incisivos con dos modelos finales en lugar de uno solo, que podría no dar resultado.

A partir de este punto, hablaremos de dos poblaciones: la **población V1** que será de los contratos que a la fecha de corte tenían un año o menos de afiliación y la **población V2** que será de los contratos que tienen más de un año de afiliación.

4.5.6 MUESTREO PARA PARTICIÓN EN BASES DE DESARROLLO Y PRUEBA

Se debe comprobar que cada uno de los modelos desarrollados cumple los supuesto y que además tenga un ajuste adecuado y una capacidad de predicción y discriminación adecuados. Con este fin, se decidió tomar la opción de hacerlo mediante la metodología de partición de la base de datos de cada población en dos grupos, el primer grupo será la llamada base de desarrollo y el segundo la base de prueba.

La base de desarrollo será usada para la construcción del modelo mientras que la base de prueba es una data completamente independiente que se usará para verificar los resultados obtenidos al elaborar el modelo. De esta manera, podremos verificar que no tengamos un modelo sobre-entrenado y que los resultados de las predicciones sean confiables.

También existen métodos de prueba como la validación cruzada, sin embargo, considerando que tenemos suficientes datos usaremos la partición Desarrollo/Prueba o Entrenamiento/-Validación.

Para la partición se usó un muestreo aleatorio estratificado donde los estratos están determinados por “sujetos censurados” y “sujetos no censurados”. De esta forma nos aseguramos de tener una proporción igual en cada base.

Se usarán 2/3 de la población para el desarrollo de cada modelo y 1/3 para pruebas, estos valores han sido seleccionados en base a la recomendación de expertos en sus trabajos. La tabla 4.3 muestra los resultados finales de la partición.

Tabla 4.3: Partición Desarrollo/Prueba de las bases de datos de las dos poblaciones propuestas por muestreo estratificado

Población	Desarrollo		Prueba	
	Censurados	No Censurados	Censurados	No Censurados
Afiliados 1 año o menos	1902	1305	951	652
Afiliados más de 1 año	3460	1252	1730	625

4.5.7 ANÁLISIS DESCRIPTIVO Y DEPURACIÓN DE LAS BASES DE DATOS POR CADA POBLACIÓN

Una vez que hemos dividido los datos en una base de desarrollo y una de prueba para cada población, regresamos al punto en el que debemos realizar un análisis descriptivo de las variables y depurar las bases. Este proceso se lo realiza únicamente para las bases de desarrollo, ya que las bases de prueba nos servirán para verificar los resultados finales y por ende no pueden ser tocadas.

4.5.7.1 Análisis descriptivo y Depuración de la Base de Datos de la Población V1

La tabla 4.4 muestra un análisis descriptivo de las variables para la base de desarrollo de la población V1. Debido a que muchas de las variables son bastante complejas de construir -y por la misma razón sus nombres no siempre pueden ser lo suficientemente descriptivos- se creó un **Diccionario de Variables** (ver anexo 7.4) a través del cual el lector puede comprender que es lo que cada variable mide o muestra.

En base al análisis descriptivo se hará una depuración de las variables por cada uno de los criterios más importantes. La tabla 4.4 nos muestra que hay muchas variables con una gran cantidad de valores perdidos. Las siguientes covariables serán eliminadas del estudio debido a que tiene más del 35 % de valores perdidos:

- ❒ PORCENTAJE_COBERTURA
- ❒ PORCENTAJE_COBERTURA_3M_PA
- ❒ PORCENTAJE_COBERTURA_6M_PA
- ❒ PORCENTAJE_COBERTURA_9M_PA
- ❒ PORCENTAJE_COBERTURA_12M_PA
- ❒ PORCENTAJE_COBERTURA_3M
- ❒ PORCENTAJE_COBERTURA_6M
- ❒ PORCENTAJE_COBERTURA_9M

Tabla 4.4: Descripción de Variables Población V1

Variable	Mínimo	Primer Cuartil	Mediana	Media	Tercer Cuartil	Máximo	Valores Perdidos
CONTRATO							
CENSURA							
TIEMPO	0	196	366	278,8	366	366	0
PRESENTADO	0	0	0	102,31	53,67	20394,31	0
PAGADO_NETO	0	0	0	70,98	32,25	13397,98	0
PORCENTAJE_COBERTURA	0	0,5426	0,697	0,6342	0,777	1,0878	1613
PRESENTADO_3M_PA	0	0	0	37,97	0	19870,88	0
PAGADO_NETO_3M_PA	0	0	0	20,44	0	13079,73	0
PORCENTAJE_COBERTURA_3M_PA	0	0,1394	0,48	0,4333	0,6715	1	2829
PRESENTADO_6M_PA	0	0	0	84,25	0	19870,88	0
PAGADO_NETO_6M_PA	0	0	0	44,67	0	13079,73	0
PORCENTAJE_COBERTURA_6M_PA	0	0,1336	0,4232	0,3941	0,6119	1	2599
PRESENTADO_9M_PA	0	0	0	119,8	0	19870,9	0
PAGADO_NETO_9M_PA	0	0	0	61,84	0	13079,73	0
PORCENTAJE_COBERTURA_9M_PA	0	0,123	0,4	0,369	0,5794	1	2510
PRESENTADO_12M_PA	0	0	0	132,6	0	19870,9	0
PAGADO_NETO_12M_PA	0	0	0	68,65	0	13079,73	0
PORCENTAJE_COBERTURA_12M_PA	0	0,121	0,3968	0,3615	0,5666	1	2501
PRESENTADO_3M	0	0	0	102,31	53,67	20394,31	0
PAGADO_NETO_3M	0	0	0	70,98	32,25	13397,98	0
PORCENTAJE_COBERTURA_3M	0	0,5802	0,7	0,6533	0,7864	1	2099
PRESENTADO_6M	0	0	0	255,3	164,5	57372,3	0
PAGADO_NETO_6M	0	0	0	182,3	102,3	51097,1	0
PORCENTAJE_COBERTURA_6M	0	0,5605	0,7	0,6472	0,7819	1,0878	1692
PRESENTADO_9M	0	0	0	335,2	227,6	57837,7	0
PAGADO_NETO_9M	0	0	0	234,2	139	51466	0
PORCENTAJE_COBERTURA_9M	0	0,5484	0,6997	0,6362	0,7791	1,0878	1622
PRESENTADO_12M	0	0	0	355,6	236,4	57837,7	0
PAGADO_NETO_12M	0	0	0	247,1	145,1	51466	0
PORCENTAJE_COBERTURA_12M	0	0,543	0,6971	0,6343	0,777	1,0878	1613
RECORTE_12M	0	0	0	13,56	0	2210,47	0
RECORTE_9M	0	0	0	12,03	0	1224,57	0
RECORTE_6M	0	0	0	8,933	0	1089	0
RECORTE_3M	0	0	0	3,676	0	1020	0
REEMBOLSOS_CERO_PAGO_3M	0	0	0	0,04646	0	3	0
REEMBOLSOS_CERO_PAGO_6M	0	0	0	0,1235	0	5	0
REEMBOLSOS_CERO_PAGO_9M	0	0	0	0,1746	0	5	0
REEMBOLSOS_CERO_PAGO_12M	0	0	0	0,1893	0	6	0
MAX_TIEMPO_OCURRENCIA_PAGO_12M	2	34	44	46,65	58	185	1613
MAX_TIEMPO_OCURRENCIA_PAGO_9M	2	34	44	46,11	58	185	1622
MAX_TIEMPO_OCURRENCIA_PAGO_6M	2	33	41	44,24	57	127	1692
MAX_TIEMPO_OCURRENCIA_PAGO_3M	2	29	36	35,13	40	70	2099
EDAD_TITULAR	18	29	35	39,21	47	106	0
GENERO_TITULAR	0	0	0	0,4766	1	1	0
NUMERO_AFILIADOS	1	1	1	1,716	2	10	0
NUMERO_AFILIADOS_MENORES	0	0	0	0,5182	1	7	0
NUMERO_AFILIADOS_5AÑOS_O_MENOS	0	0	0	0,2205	0	3	0
NUMERO_AFILIADOS_2AÑOS_O_MENOS	0	0	0	0,1188	0	2	0
INDIVIDUAL_FAMILIAR	1	1	1	1,383	2	2	0
CONYUGE	0	0	0	0,1693	0	1	0
TIEMPO_AFILIACION	3	92	169	179,1	261,5	364	0
PRODUCTO_MH	0	0	0	0,2934	1	1	0
TOPE_PLAN	15000	30000	50000	51241	80000	150000	0
TITULAR_SIN_BENEFICIOS	0	0	0	0,07764	0	1	0
PRESENTADO_CRONICO_12M	0	0	0	102,15	16,34	56868,6	0
PAGADO_NETO_CRONICO_12M	0	0	0	76,22	6	50701,73	0
PORCENTAJE_COBERTURA_CRONICO_12M	0	0,6053	0,7172	0,6732	0,8012	1	2365
PRESENTADO_CPC_12M	0	0	0	248,4	105,7	57595	0
PAGADO_NETO_CPC_12M	0	0	0	178,1	59,5	51276,7	0
PORCENTAJE_COBERTURA_CPC_12M	0	0,5724	0,7017	0,6535	0,7974	1,0878	1896
INCLUSIONES_3M	0	0	0	0,009354	0	3	0
INCLUSIONES_6M	0	0	0	0,02494	0	4	0
INCLUSIONES_9M	0	0	0	0,02962	0	7	0
INCLUSIONES_12M	0	0	0	0,02962	0	7	0
EXCLUSIONES_3M	0	0	0	0,007484	0	4	0
EXCLUSIONES_6M	0	0	0	0,0159	0	4	0
EXCLUSIONES_9M	0	0	0	0,02089	0	7	0
EXCLUSIONES_12M	0	0	0	0,02183	0	7	0
CAMBIO_PLAN_3M	0	0	0	0,1706	0	1	0
CAMBIO_PLAN_6M	0	0	0	0,2404	0	1	0
CAMBIO_PLAN_9M	0	0	0	0,4079	1	1	0
CAMBIO_PLAN_12M	0	0	0	0,4833	1	1	0
MOVIMIENTOS_3M	0	0	0	0,3673	0	9	0
MOVIMIENTOS_6M	0	0	0	0,575	1	13	0
MOVIMIENTOS_9M	0	0	0	0,9245	1	19	0
MOVIMIENTOS_12M	0	0	1	1,091	2	19	0
MORA_PROM_12M	0	0	0	1,874	0,75	47,75	2
MORA_MAX_12M	0	0	0	6,652	4	169	2
MORA_PROM_6M	0	0	0	2,0753	0,6667	49,8	15
MORA_MAX_6M	0	0	0	6,215	3	169	15
MORA_PROM_3M	0	0	0	2,387	0	61,5	42
MORA_MAX_3M	0	0	0	4,817	0	77	42
INCREMENTO_RELATIVO_2018	-0,2638	0	0,0194	0,1573	0,0458	2,3038	2978
INCREMENTO_ABSOLUTO_2018	-23,1	0	1,5	11,37	3	269,64	2978
DIAS_FIN_CONTRATO	0	102,5	195	184,9	272	361	0
NUEVO	0	0	1	0,5179	1	1	0
USO	0	0	0	0,4836	1	1	0
USO_6M	0	0	0	0,4615	1	1	0
USO_12M	0	0	0	0,4836	1	1	0
USO_HOSPITALARIO	0	0	0	0,005301	0	1	0
USO_AMB_EMR_HSD	0	0	0	0,2133	0	1	0
TIEMPO_SIN_USO	3	50	81	106,4	138	363	0
SINIESTRALIDAD_12M	0	0	0	0,257	0,1509	106,8523	0
SINIESTRALIDAD_6M	0	0	0	0,2026	0,1071	106,0864	0
CANAL	0	1	1	0,8566	1	1	0
PRIMA_PE	17,86	39,92	60,47	73,06	85,71	387,53	0
INCREMENTO_ABS	0	0	0	0	0	0	0
INCREMENTO_REL	0	0	0	0	0	0	0
PAGOS_EFECTIVO_12M	0	0	0	0,008514	0	1	2
PAGOS_DEBITO_BANCARIO_12M	0	0	1	0,6373	1	1	2
PAGOS_TARJETA_CREDITO_12M	0	0	0	0,3538	1	1	2
PAGOS_EFECTIVO_6M	0	0	0	0,007772	0	1	15
PAGOS_DEBITO_BANCARIO_6M	0	0	1	0,6368	1	1	15
PAGOS_TARJETA_CREDITO_6M	0	0	0	0,3551	1	1	15
PAGOS_EFECTIVO_3M	0	0	0	0,00516	0	1	42
PAGOS_DEBITO_BANCARIO_3M	0	0	1	0,6391	1	1	42
PAGOS_TARJETA_CREDITO_3M	0	0	0	0,3554	1	1	42

- ❒ PORCENTAJE_COBERTURA_12M
- ❒ MAX_TIEMPO_OCURRENCIA_PAGO_12M
- ❒ MAX_TIEMPO_OCURRENCIA_PAGO_9M
- ❒ MAX_TIEMPO_OCURRENCIA_PAGO_6M
- ❒ MAX_TIEMPO_OCURRENCIA_PAGO_3M
- ❒ PORCENTAJE_COBERTURA_CRONICO_12M
- ❒ PORCENTAJE_COBERTURA_CPC_12M
- ❒ INCREMENTO_RELATIVO_2018
- ❒ INCREMENTO_ABSOLUTO_2018

Existen variables con un solo valor, por lo que carecen de significado y por tanto fueron eliminadas:

- ❒ INCREMENTO_ABS
- ❒ INCREMENTO_REL

Además, se eliminaron a los sujetos que tienen valores perdidos en las variables:

- ❒ MORA_PROM_12M
- ❒ MORA_MAX_12M
- ❒ MORA_PROM_6M
- ❒ MORA_MAX_6M
- ❒ MORA_PROM_3M
- ❒ MORA_MAX_3M

Esta decisión implica una perdida mínima de datos (1,3 %) y así podemos conservar estas variables que han demostrado ser valiosas para la predicción (ver figura 7.9). Además, si alguna de las variables llega a entrar al modelo y en la base de pruebas existen valores perdidos existen varias estrategias para lidiar con ese problema.

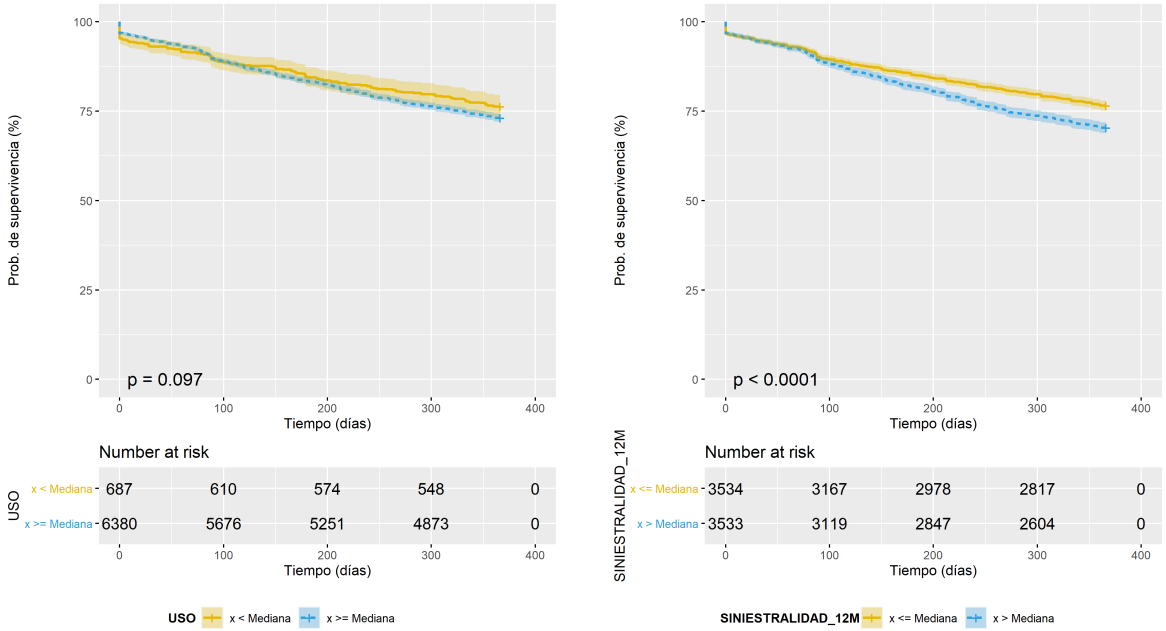


Figura 4.3: Población V1 - Ejemplos de estimador Kaplan y Meier y prueba Log-rank de dos variable de la base de desarrollo partidas dos grupos por su respectiva mediana

Para poder verificar el poder predictivo de cada variable se usa prueba log-rank. Para esto se divide a los sujetos en dos grupos por la mediana de la variable, de manera que el primer grupo estará conformado por los sujetos cuyo valor en esa variable sea 'menor o igual'[7] a la mediana de esa variable y el segundo grupo incluye a todos los demás. Luego, se compara las funciones de riesgo de cada grupo gráficamente y con la prueba log-rank (igual que antes, se usó la función de peso $W(t) = 1$).

La figura 4.3 muestra dos ejemplos de este proceso, a la izquierda de esta figura está la variable "USO" que al ser dividida en dos grupos no demostró tener buena capacidad de discriminación, ya que el p-valor de la prueba log-rank[8] está por encima de 0,05 y no es suficiente para rechazar la hipótesis nula, por lo que sería eliminada del proceso. A la derecha está la variable "SINIESTRALIDAD_12M" que, por el contrario, demostró tener la capacidad de discriminar a los sujetos. Las gráficas de los estimadores de Kaplan y Meier y las pruebas log-rank para todas las variables en la población V1 pueden verse en el anexo 7.5.

[7]Para algunas variables binarias se cambió esta definición a 'menor'ya que, como la mediana solo puede tener dos valores posibles, ocurría que uno de los dos grupos absorbía a todos los sujetos y en tal caso la prueba era estéril.

[8]El p-valor de la prueba se encuentra la parte inferior izquierda del sistema de coordenadas de la gráfica del estimador de Kaplan y Meier.

Para poder discriminar mejor las variables incluiremos un análisis de correlación con el que compararemos cada variable entre sí y, más importante, cada predictor contra la censura, es decir, que veremos que tanto está correlacionada cada covariable con el hecho de que a los sujetos les ocurra o no el evento de interés en un periodo de un año. Para este análisis usaremos la correlación de Spearman o rho de Spearman debido a que se obtienen mejores resultados al comparar variables continuas con variables discretas, como la censura, y variables discretas entre sí (ver anexo 7.1). Lo haremos de igual manera para la población V2. La tabla 4.5 muestra los resultados de este cálculo, que nos dará una ayuda extra para eliminar variables poco o nada útiles.

Eliminamos datos atípicos mediante la distancia de Mahalanobis (ver anexo 7.2), para este cálculo se decidió retirar todas las variables que no muestren capacidad de discriminación o que no tengan correlación con la censura ya que puede decirse que no influyen en la ocurrencia del evento de interés. Recordemos que al eliminar datos atípicos se buscan cumplir con los supuestos del modelo de Cox de que los sujetos no influyen en la estimación de los coeficientes ni en la estimación del modelo, de manera que no sería adecuado considerar atípico a un dato en base a una variable que no influirá en el modelo.

En base a las pruebas log-rank mostradas en las figuras del anexo 7.5 y, en segundo lugar, a la tabla 4.5 quitamos las variables con poca capacidad de discriminación y no correlacionadas, y nos quedamos con 21 de 85 variables restantes. Estas variables fueron:

- ❒ PRESENTADO_6M
- ❒ PAGADO_NETO_6M
- ❒ PAGADO_NETO_12M
- ❒ EDAD_TITULAR
- ❒ NUMERO_AFILIADOS
- ❒ INDIVIDUAL_FAMILIAR
- ❒ TOPE_PLAN
- ❒ CAMBIO_PLAN_9M
- ❒ CAMBIO_PLAN_12M
- ❒ MOVIMIENTOS_9M

Tabla 4.5: Correlación de Spearman o rho de Spearman de cada covariable comparada con la Censura en la población V1

Variable	Estadístico	P-valor
CENSURA		
PRESENTADO	-0.058	0.001
PAGADO_NETO	-0.059	0.001
PRESENTADO_3M_PA	-0.061	0.001
PAGADO_NETO_3M_PA	-0.060	0.001
PRESENTADO_6M_PA	-0.065	0.0003
PAGADO_NETO_6M_PA	-0.059	0.001
PRESENTADO_9M_PA	-0.060	0.001
PAGADO_NETO_9M_PA	-0.055	0.002
PRESENTADO_12M_PA	-0.060	0.001
PAGADO_NETO_12M_PA	-0.058	0.001
PRESENTADO_3M	-0.058	0.001
PAGADO_NETO_3M	-0.059	0.001
PRESENTADO_6M	-0.047	0.008
PAGADO_NETO_6M	-0.047	0.009
PRESENTADO_9M	-0.042	0.017
PAGADO_NETO_9M	-0.040	0.026
PRESENTADO_12M	-0.044	0.013
PAGADO_NETO_12M	-0.042	0.018
RECORTE_12M	-0.076	0.00002
RECORTE_9M	-0.076	0.00002
RECORTE_6M	-0.074	0.00003
RECORTE_3M	-0.087	0.00000
REEMBOLSOS_CERO_PAGO_3M	-0.029	0.097
REEMBOLSOS_CERO_PAGO_6M	-0.061	0.001
REEMBOLSOS_CERO_PAGO_9M	-0.073	0.00004
REEMBOLSOS_CERO_PAGO_12M	-0.070	0.0001
EDAD_TITULAR	-0.082	0.00000
GENERO_TITULAR	-0.035	0.048
NUMERO_AFILIADOS	0.058	0.001
NUMERO_AFILIADOS_MENORES	0.084	0.00000
NUMERO_AFILIADOS_5AÑOS_O_MENOS	0.062	0.0005
NUMERO_AFILIADOS_2AÑOS_O_MENOS	0.055	0.002
INDIVIDUAL_FAMILIAR	0.062	0.001
CONYUGE	-0.035	0.052
TIEMPO_AFILIACION	-0.019	0.283
PRODUCTO_MH	-0.050	0.005
TOPE_PLAN	-0.085	0.00000
TITULAR_SIN_BENEFICIOS	-0.005	0.792
PRESENTADO_CRONICO_12M	-0.044	0.013
PAGADO_NETO_CRONICO_12M	-0.038	0.032
PRESENTADO_CPC_12M	-0.043	0.015
PAGADO_NETO_CPC_12M	-0.040	0.026
INCLUSIONES_3M	0.009	0.606
INCLUSIONES_6M	0.009	0.620
INCLUSIONES_9M	0.007	0.714
INCLUSIONES_12M	0.007	0.714
EXCLUSIONES_3M	-0.005	0.798
EXCLUSIONES_6M	0.008	0.643
EXCLUSIONES_9M	0.026	0.141
EXCLUSIONES_12M	0.027	0.135
CAMBIO_PLAN_3M	0.025	0.158
CAMBIO_PLAN_6M	0.029	0.107
CAMBIO_PLAN_9M	0.056	0.002
CAMBIO_PLAN_12M	0.057	0.001
MOVIMIENTOS_3M	0.025	0.156
MOVIMIENTOS_6M	0.032	0.073
MOVIMIENTOS_9M	0.065	0.0002
MOVIMIENTOS_12M	0.064	0.0003
MORA_PROM_12M	0.240	0
MORA_MAX_12M	0.235	0
MORA_PROM_6M	0.248	0
MORA_MAX_6M	0.244	0
MORA_PROM_3M	0.257	0
MORA_MAX_3M	0.255	0
DIAS_FIN_CONTRATO	0.019	0.282
NUEVO	0.015	0.407
USO	-0.045	0.011
USO_6M	-0.047	0.009
USO_12M	-0.045	0.011
USO_HOSPITALARIO	0.001	0.963
USO_AMB_EMR_HSD	-0.053	0.003
TIEMPO_SIN_USO	0.045	0.012
SINIESTRALIDAD_12M	-0.027	0.131
SINIESTRALIDAD_6M	-0.031	0.079
CANAL	0.114	0
PRIMA_PE	-0.094	0.00000
PAGOS_EFECTIVO_12M	0.008	0.643
PAGOS_DEBITO_BANCARIO_12M	0.145	0
PAGOS_TARJETA_CREDITO_12M	-0.147	0
PAGOS_EFECTIVO_6M	-0.004	0.815
PAGOS_DEBITO_BANCARIO_6M	0.147	0
PAGOS_TARJETA_CREDITO_6M	-0.146	0
PAGOS_EFECTIVO_3M	0.010	0.590
PAGOS_DEBITO_BANCARIO_3M	0.144	0
PAGOS_TARJETA_CREDITO_3M	-0.145	0

- ❒ MOVIMIENTOS_12M
- ❒ USO_6M
- ❒ TIEMPO_SIN_USO
- ❒ CANAL
- ❒ PRIMA_PE
- ❒ PAGOS_DEBITO_BANCARIO_12M
- ❒ PAGOS_TARJETA_CREDITO_12M
- ❒ PAGOS_DEBITO_BANCARIO_6M
- ❒ PAGOS_TARJETA_CREDITO_6M
- ❒ PAGOS_DEBITO_BANCARIO_3M
- ❒ PAGOS_TARJETA_CREDITO_3M

La distancia de mahalanobis sigue aproximadamente una χ^2 con n grados de libertad, donde n es el número de variables usado, en nuestro caso 21. Se tomó cómo dato atípico a todo aquel que tenga un p-valor menor a 0,001. En función de estas variables se determinó como datos atípicos a 143 de 3165 sujetos, es decir, el 4,5 % de los datos. La tabla siguiente muestra cómo se repartieron estos datos en función de la censura.

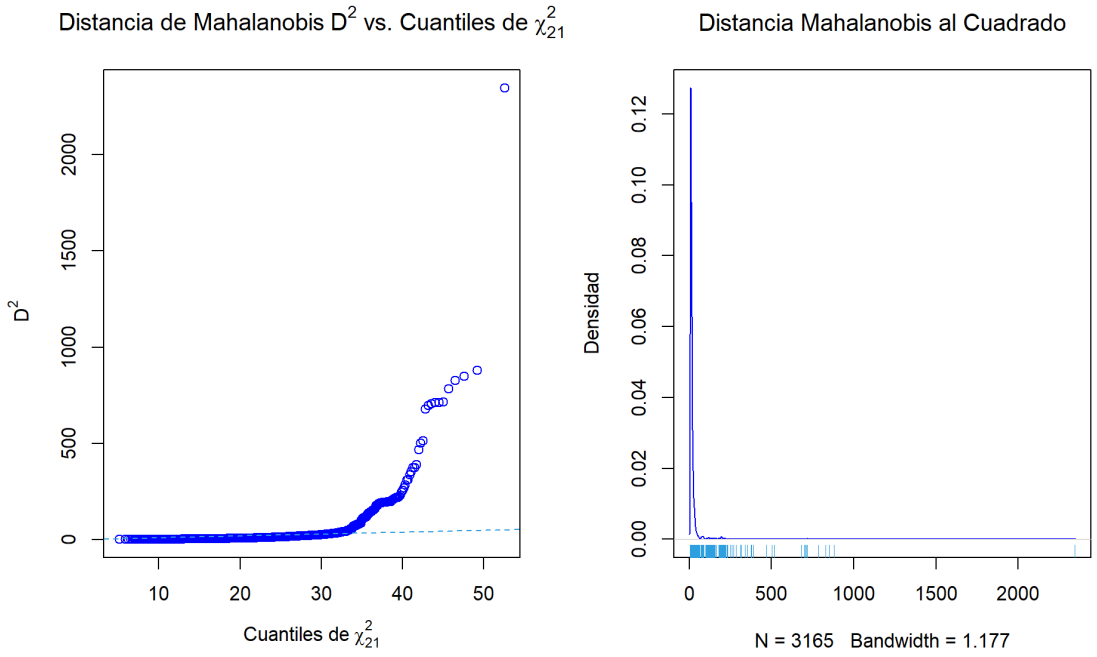
Tabla 4.6: Tabla de contingencia de Censura contra datos atípicos V1

	No Atípico	**Atípico**
Censurado	1801	78
No censurado	1221	65

Mediante una prueba de independencia chi-cuadrado aplicada a la tabla 4.6 podemos determinar si los datos atípicos están distribuidos por igual en los grupo de datos censurados y no censurados o están más concentrados en alguno de ellos. El resultado de la prueba fue el siguiente:

Código de R 4.2: Salida de R prueba chi-cuadrado de la tabla 4.6

```
1         Pearson's Chi-squared test with Yates' continuity correction
-
-  data:  tabla.atipicosV1
-  X-squared = 1.2422, df = 1, p-value = 0.265
```



(a) Distancia de Mahalanobis al cuadrado contra cuantiles de la χ^2_{21} correspondiente, en la población V1

(b) función de densidad de la distancia de Mahalanobis para la población V1

Figura 4.4: Distancia de Mahalanobis para la población V1

Cómo el p-valor de la prueba es mayor a 0,05 no rechazamos la hipótesis nula, entonces podemos decir que los datos se distribuyen entre ambos grupos.

La figura 4.4a nos muestra cómo se distribuyen los cuantiles de los datos, en caso de tener una distribución perfecta, formarían una línea recta a 45 grados como la que esta representada con una línea segmentada. En la figura 4.4b podemos ver la densidad de la distancia de Mahalanobis y donde están mayoritariamente acumulados los datos.

4.5.7.2 Análisis descriptivo y Depuración de la Base de Datos V2

La tabla 4.7 muestra un análisis descriptivo de las variables para la base de desarrollo de la población V2. Como ya se mencionó en el apartado anterior, se creó un **Diccionario de Variables** (ver anexo 7.4) como ayuda al lector. Esta población corresponde a los contratos que tienen más de un año en la compañía, por lo que no existen contratos nuevos en esta población, de manera que la variable “NUEVO” se elimina de antemano. En base al análisis descriptivo se hará una depuración de las variables por cada uno de los criterios más importantes. La tabla 4.7 nos muestra que, al igual que en la población anterior, hay

Tabla 4.7: Descripción de Variables Población V2

Variable	Mínimo	Primer Cuartil	Mediana	Media	Tercer Cuartil	Máximo	Valores Perdidos
CONTRATO							
CENSURA							
TIEMPO	0	333	366	307,6	366	366	0
PRESENTADO	0	0	0	180,9	106,4	17576,5	0
PAGADO_NETO	0	0	0	127,25	64,28	14503,79	0
PORCENTAJE_COBERTURA	0	0,5948	0,7003	0,6679	0,7732	1	447
PRESENTADO_3M_PA	0	0	0	74,25	0	16786,38	0
PAGADO_NETO_3M_PA	0	0	0	42,11	0	12420,83	0
PORCENTAJE_COBERTURA_3M_PA	0	0,178	0,501	0,447	0,7	1	3835
PRESENTADO_6M_PA	0	0	0	175,5	100	16836,4	0
PAGADO_NETO_6M_PA	0	0	0	99,04	26,06	12420,83	0
PORCENTAJE_COBERTURA_6M_PA	0	0,243	0,49	0,446	0,667	1	3204
PRESENTADO_9M_PA	0	0	0	288,7	220	21634,3	0
PAGADO_NETO_9M_PA	0	0	0	159,68	82,53	12420,83	0
PORCENTAJE_COBERTURA_9M_PA	0	0,2446	0,4727	0,4294	0,6275	1	2790
PRESENTADO_12M_PA	0	0	0	417,8	348,6	37291,1	0
PAGADO_NETO_12M_PA	0	0	0	236,4	162	25588,9	0
PORCENTAJE_COBERTURA_12M_PA	0	0,28	0,4993	0,45	0,6369	1	2468
PRESENTADO_3M	0	0	0	180,9	106,4	17576,5	0
PAGADO_NETO_3M	0	0	0	127,25	64,28	14503,79	0
PORCENTAJE_COBERTURA_3M	0	0,5724	0,7	0,6389	0,7742	1	2536
PRESENTADO_6M	0	0	103,5	528,8	429,8	41479,9	0
PAGADO_NETO_6M	0	0	60,8	382,4	280,1	36648,7	0
PORCENTAJE_COBERTURA_6M	0	0,588	0,7064	0,6561	0,7839	1	1505
PRESENTADO_9M	0	26	236,8	885	794,3	41619,3	0
PAGADO_NETO_9M	0	9,95	145,71	641,65	537,02	36766,03	0
PORCENTAJE_COBERTURA_9M	0	0,5879	0,7102	0,6592	0,7871	1	1068
PRESENTADO_12M	0	65,44	394,94	1266,07	1232,34	76859,89	0
PAGADO_NETO_12M	0	34,44	250,88	924,51	840,53	59420,99	0
PORCENTAJE_COBERTURA_12M	0	0,5985	0,7106	0,6689	0,7878	1	822
RECORTE_12M	0	0	0	39,9	25,55	2984,75	0
RECORTE_9M	0	0	0	29,38	15	2513,94	0
RECORTE_6M	0	0	0	18,27	5,2	1506,62	0
RECORTE_3M	0	0	0	6,835	0	913,75	0
REEMBOLSOS_CERO_PAGO_3M	0	0	0	0,08701	0	5	0
REEMBOLSOS_CERO_PAGO_6M	0	0	0	0,2203	0	7	0
REEMBOLSOS_CERO_PAGO_9M	0	0	0	0,3758	0	10	0
REEMBOLSOS_CERO_PAGO_12M	0	0	0	0,4798	1	10	0
MAX_TIEMPO_OCURRENCIA_PAGO_12M	2	38	53	53,52	65	273	822
MAX_TIEMPO_OCURRENCIA_PAGO_9M	2	37	49	49,96	62	232	1068
MAX_TIEMPO_OCURRENCIA_PAGO_6M	0	35	44	45,6	58	133	1505
MAX_TIEMPO_OCURRENCIA_PAGO_3M	2	30	36	36,39	44	83	2536
EDAD_TITULAR	18	33	40	42,59	52	96	0
GENERO_TITULAR	0	0	0	0,4619	1	1	0
NUMERO_AFILIADOS	1	1	1	1,684	2	10	0
NUMERO_AFILIADOS_MENORES	0	0	0	0,545	1	6	0
NUMERO_AFILIADOS_5AÑOS_O_MENOS	0	0	0	0,2143	0	3	0
NUMERO_AFILIADOS_2AÑOS_O_MENOS	0	0	0	0,08383	0	2	0
INDIVIDUAL_FAMILIAR	1	1	1	1,365	2	2	0
CONYUGE	0	0	0	0,195	0	1	0
TIEMPO_AFILIACION	368	610	899	1303	1614	4014	0
PRODUCTO_MH	0	0	1	0,5458	1	1	0
TOPE_PLAN	15000	30000	50000	54552	80000	150000	0
TITULAR_SIN_BENEFICIOS	0	0	0	0,2146	0	1	0
PRESENTADO_CRONICO_12M	0	0	29,65	436,97	260,35	48677,08	0
PAGADO_NETO_CRONICO_12M	0	0	17,66	316,18	169,22	41376,83	0
PORCENTAJE_COBERTURA_CRONICO_12M	0	0,6182	0,7261	0,6796	0,8077	1	2053
PRESENTADO_CPC_12M	0	0	170,1	901,4	715,4	74825,7	0
PAGADO_NETO_CPC_12M	0	0	102,4	667,3	477,8	58156,8	0
PORCENTAJE_COBERTURA_CPC_12M	0	0,6072	0,7249	0,6771	0,8023	1	1253
INCLUSIONES_3M	0	0	0	0,01167	0	3	0
INCLUSIONES_6M	0	0	0	0,02271	0	3	0
INCLUSIONES_9M	0	0	0	0,03587	0	4	0
INCLUSIONES_12M	0	0	0	0,04648	0	5	0
EXCLUSIONES_3M	0	0	0	0,01167	0	6	0
EXCLUSIONES_6M	0	0	0	0,02462	0	6	0
EXCLUSIONES_9M	0	0	0	0,04966	0	8	0
EXCLUSIONES_12M	0	0	0	0,06473	0	8	0
CAMBIO_PLAN_3M	0	0	0	0,01082	0	1	0
CAMBIO_PLAN_6M	0	0	0	0,02589	0	1	0
CAMBIO_PLAN_9M	0	0	0	0,05518	0	1	0
CAMBIO_PLAN_12M	0	0	0	0,0677	0	1	0
MOVIMIENTOS_3M	0	0	0	0,1231	0	26	0
MOVIMIENTOS_6M	0	0	0	0,229	0	26	0
MOVIMIENTOS_9M	0	0	0	0,4119	0	27	0
MOVIMIENTOS_12M	0	0	0	0,5357	0	27	0
MORA_PROM_12M	0	0	0	2,135	2,583	125	0
MORA_MAX_12M	0	0	0	8,489	6	138	0
MORA_PROM_6M	0	0	0	2,286	2,333	125	15
MORA_MAX_6M	0	0	0	6,294	4	138	15
MORA_PROM_3M	0	0	0	2,274	1,356	64	30
MORA_MAX_3M	0	0	0	4,281	3	77	30
INCREMENTO_RELATIVO_2018	-0,45146	0	0,07109	0,16876	0,1686	4,11789	7
INCREMENTO_ABSOLUTO_2018	-76,73	0	4,71	18	19,63	722,35	7
DIAS_FIN_CONTRATO	13	102	150	172,3	242	361	0
USO	0	1	1	0,8983	1	1	0
USO_6M	0	0	1	0,6691	1	1	0
USO_12M	0	1	1	0,8205	1	1	0
USO_HOSPITALARIO	0	0	0	0,009338	0	1	0
USO_AMB_EMR_HSD	0	0	0	0,2685	1	1	0
TIEMPO_SIN_USO	8	55	98	210,5	246	1065	0
SINIESTRALIDAD_12M	0	0,01533	0,09927	0,30049	0,28451	15,32466	0
SINIESTRALIDAD_6M	0	0	0,02346	0,12862	0,09665	12,29723	0
CANAL	0	1	1	0,8877	1	1	0
PRIMA_PE	0,08	48,08	69,54	82,83	100,35	387,53	0
INCREMENTO_ABS	-4493,42	0	4,39	16,69	19,14	4259,34	0
INCREMENTO_REL	-0,68874	0	0,06446	0,17555	0,15568	28,70162	0
PAGOS_EFECTIVO_12M	0	0	0	0,01295	0	1	0
PAGOS_DEBITO_BANCARIO_12M	0	0	1	0,5754	1	1	0
PAGOS_TARJETA_CREDITO_12M	0	0	0	0,4117	1	1	0
PAGOS_EFECTIVO_6M	0	0	0	0,01134	0	1	15
PAGOS_DEBITO_BANCARIO_6M	0	0	1	0,5761	1	1	15
PAGOS_TARJETA_CREDITO_6M	0	0	0	0,4126	1	1	15
PAGOS_EFECTIVO_3M	0	0	0	0,01004	0	1	30
PAGOS_DEBITO_BANCARIO_3M	0	0	1	0,5761	1	1	30
PAGOS_TARJETA_CREDITO_3M	0	0	0	0,4139	1	1	30

muchas variables con una gran cantidad de valores perdidos. Las siguientes covariables serán eliminadas del estudio debido a que tiene más del 35 % de valores perdidos:

- ❒ PORCENTAJE_COBERTURA_3M_PA
- ❒ PORCENTAJE_COBERTURA_6M_PA
- ❒ PORCENTAJE_COBERTURA_9M_PA
- ❒ PORCENTAJE_COBERTURA_12M_PA
- ❒ PORCENTAJE_COBERTURA_3M
- ❒ PORCENTAJE_COBERTURA_6M
- ❒ PORCENTAJE_COBERTURA_9M
- ❒ MAX_TIEMPO_OCURRENCIA_PAGO_9M
- ❒ MAX_TIEMPO_OCURRENCIA_PAGO_6M
- ❒ MAX_TIEMPO_OCURRENCIA_PAGO_3M
- ❒ PORCENTAJE_COBERTURA_CRONICO_12M
- ❒ PORCENTAJE_COBERTURA_CPC_12M

En base al análisis descriptivo, se eliminaron a los sujetos que tienen valores perdidos en las variables:

- ❒ MORA_PROM_12M
- ❒ MORA_MAX_12M
- ❒ MORA_PROM_6M
- ❒ MORA_MAX_6M
- ❒ MORA_PROM_3M
- ❒ MORA_MAX_3M
- ❒ INCREMENTO_RELATIVO_2018
- ❒ INCREMENTO_ABSOLUTO_2018

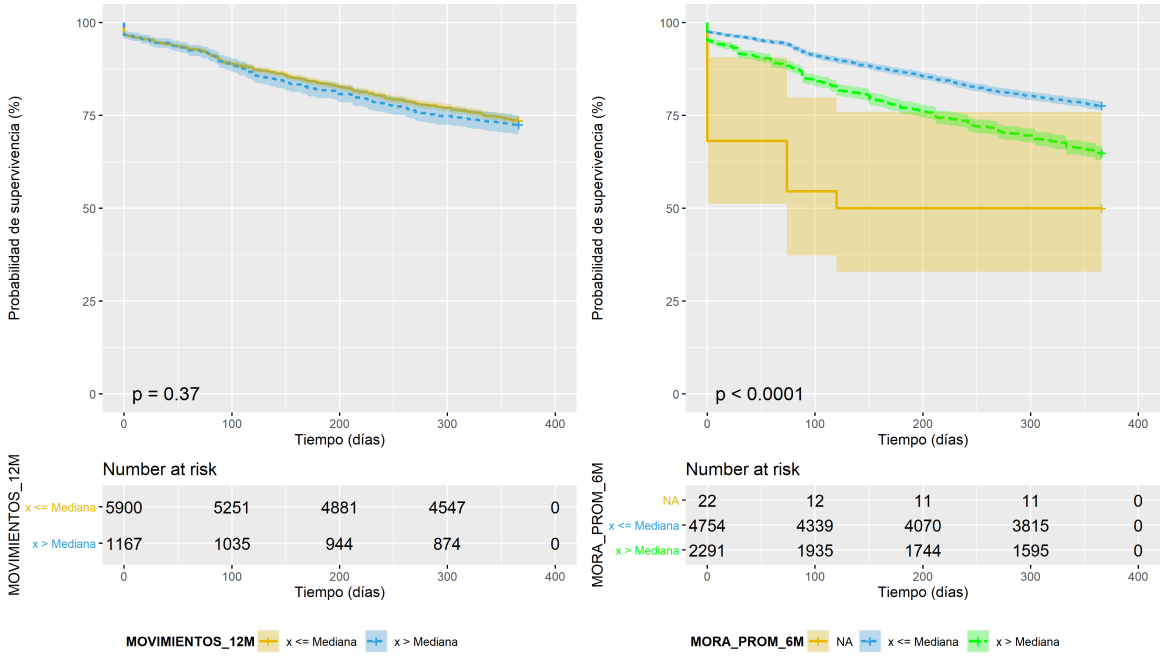


Figura 4.5: Población V2 - Ejemplos de estimador Kaplan y Meier y prueba Log-rank de dos variable de la base de desarrollo partidas dos grupos por su respectiva mediana

Esta decisión implica una perdida mínima de datos (0,7 %) pero a cambio nos permite conservar estas variables que han demostrado ser valiosas para la predicción (ver figura 7.21). Como ya se dijo antes, en caso de tener valores perdidos en la base de prueba, existen varias estrategias para lidiar con el problema.

Idénticamente que en el apartado anterior, verificaremos en la capacidad de discriminación de las covariables con la prueba log-rank. Para esto se divide a los sujetos en dos grupos por la mediana de la variable, de manera que el primer grupo estará conformado por los sujetos cuyo valor en esa variable sea 'menor o igual'[9] a la mediana de esa variable y el segundo grupo incluye a todos los demás. Luego, se compara las funciones de riesgo de cada grupo gráficamente y con la prueba log-rank (siguiendo el mismo criterio anterior, se usó la función de peso $W(t) = 1$). La figura 4.5 muestra dos ejemplos de este proceso. A la izquierda de esta figura está la variable "MOVIMIENTOS_12M" que al ser dividida en dos grupos no demostró tener buena capacidad de discriminación, ya que el p-valor de la prueba log-rank[10] está por encima de 0,05 y no es suficiente para rechazar la hipótesis nula,

[9]Tal como ocurrió con la población anterior, a algunas variables binarias se cambió esta definición a 'menor'debido a que, como la mediana puede tener solo dos valores posibles, ocurría que uno de los dos grupos absorbía a todos los sujetos y en tal caso la prueba era inútil.

[10]El p-valor de la prueba se encuentra en la parte inferior izquierda del sistema de coordenadas de la gráfica del estimador de Kaplan y Meier.

por lo que sería eliminada del proceso. A la derecha está la variable “MORA_PROM_6M” que, por el contrario, demostró tener la capacidad de discriminar a los sujetos y debemos acotar que esta variable tiene tres grupos, lo que se debe a que uno de ellos, nombrado “NA”, contiene a los sujetos que tienen valores perdidos y fue necesario incluirlo para poder usar esta información al momento de elegir una estrategia para afrontar el problema de los valores perdidos en la base de prueba. Las gráficas de los estimadores de Kaplan y Meier y las pruebas log-rank para todas las variables en la población V2 pueden verse en el anexo 7.6.

De la misma manera que con la población V1, usaremos la correlación de Spearman o rho de Spearman[11] para calcular la correlación de las variables con la censura. La tabla 4.8 muestra las correlaciones de cada predictor con la censura. Esto nos ayudará más tarde a eliminar variables poco o nada útiles. Eliminamos datos atípicos mediante la distancia de Mahalanobis (ver anexo 7.2). De la misma manera que lo hicimos con la población V1, se decidió retirar todas las variables que no muestren capacidad de discriminación o que no tengan correlación con la censura ya que puede decirse que no influyen en la ocurrencia del evento de interés. Recordemos que al eliminar datos atípicos se buscan cumplir con los supuestos del modelo de Cox de que los sujetos no influyen en la estimación de los coeficientes ni en la estimación del modelo, de manera que no sería adecuado considerar atípico a un dato en base a una variable que no influirá en el modelo.

En base a las pruebas log-rank mostradas en las figuras del anexo 7.6 y, en segundo lugar, a la tabla 4.8 quitamos las variables con poca capacidad de discriminación y no correlacionadas, y nos quedamos con 13 de 91 variables restantes. Estas variables fueron:

- ❒ EDAD_TITULAR
- ❒ NUMERO_AFILIADOS_MENORES
- ❒ TIEMPO_AFILIACION
- ❒ PRODUCTO_MH
- ❒ MORA_PROM_12M
- ❒ MORA_MAX_12M

[11] Su uso se justifica en el apartado anterior y también en el anexo 7.1.

Tabla 4.8: Correlación de Spearman o rho de Spearman de cada covariable comparada con la Censura en la población V2

Variable	Estadístico	P-valor
CENSURA		
PRESENTADO	-0.003	0.812
PAGADO_NETO	-0.003	0.857
PRESENTADO_3M_PA	-0.008	0.594
PAGADO_NETO_3M_PA	-0.004	0.774
PRESENTADO_6M_PA	-0.026	0.076
PAGADO_NETO_6M_PA	-0.026	0.076
PRESENTADO_9M_PA	-0.021	0.154
PAGADO_NETO_9M_PA	-0.020	0.181
PRESENTADO_12M_PA	-0.021	0.158
PAGADO_NETO_12M_PA	-0.017	0.242
PRESENTADO_3M	-0.003	0.812
PAGADO_NETO_3M	-0.003	0.857
PRESENTADO_6M	-0.002	0.883
PAGADO_NETO_6M	0.0005	0.974
PRESENTADO_9M	0.012	0.401
PAGADO_NETO_9M	0.015	0.313
PRESENTADO_12M	0.021	0.157
PAGADO_NETO_12M	0.024	0.108
RECORTE_12M	0.001	0.967
RECORTE_9M	-0.002	0.883
RECORTE_6M	-0.004	0.805
RECORTE_3M	0.007	0.632
REEMBOLSOS_CERO_PAGO_3M	-0.007	0.618
REEMBOLSOS_CERO_PAGO_6M	-0.020	0.168
REEMBOLSOS_CERO_PAGO_9M	-0.016	0.283
REEMBOLSOS_CERO_PAGO_12M	-0.017	0.253
EDAD_TITULAR	-0.122	0
GENERO_TITULAR	-0.035	0.016
NUMERO_AFILIADOS	0.008	0.606
NUMERO_AFILIADOS_MENORES	0.079	0.00000
NUMERO_AFILIADOS_5AÑOS_O_MENOS	0.070	0.00000
NUMERO_AFILIADOS_2AÑOS_O_MENOS	0.073	0.00000
INDIVIDUAL_FAMILIAR	0.013	0.369
CONYUGE	-0.026	0.076
TIEMPO_AFILIACION	-0.137	0
PRODUCTO_MH	-0.062	0.00002
TOPE_PLAN	-0.003	0.832
TITULAR_SIN_BENEFICIOS	0.067	0.00000
PRESENTADO_CRONICO_12M	-0.028	0.059
PAGADO_NETO_CRONICO_12M	-0.023	0.122
PRESENTADO_CPC_12M	-0.023	0.122
PAGADO_NETO_CPC_12M	-0.017	0.235
INCLUSIONES_3M	0.034	0.022
INCLUSIONES_6M	0.016	0.268
INCLUSIONES_9M	0.019	0.185
INCLUSIONES_12M	0.029	0.049
EXCLUSIONES_3M	-0.019	0.190
EXCLUSIONES_6M	-0.022	0.130
EXCLUSIONES_9M	-0.013	0.365
EXCLUSIONES_12M	-0.002	0.897
CAMBIO_PLAN_3M	0.049	0.001
CAMBIO_PLAN_6M	0.026	0.073
CAMBIO_PLAN_9M	0.021	0.151
CAMBIO_PLAN_12M	0.027	0.062
MOVIMIENTOS_3M	0.0001	0.992
MOVIMIENTOS_6M	-0.003	0.815
MOVIMIENTOS_9M	0.006	0.673
MOVIMIENTOS_12M	0.019	0.202
MORA_PROM_12M	0.145	0
MORA_MAX_12M	0.157	0
MORA_PROM_6M	0.154	0
MORA_MAX_6M	0.161	0
MORA_PROM_3M	0.153	0
MORA_MAX_3M	0.156	0
INCREMENTO_RELATIVO_2018	0.075	0.00000
INCREMENTO_ABSOLUTO_2018	0.053	0.0003
DIAS_FIN_CONTRATO	0.048	0.001
USO	0.016	0.278
USO_6M	0.008	0.568
USO_12M	0.015	0.293
USO_HOSPITALARIO	0.037	0.011
USO_AMB_EMR_HSD	-0.019	0.184
TIEMPO_SIN_USO	-0.007	0.616
SINIESTRALIDAD_12M	0.076	0.00000
SINIESTRALIDAD_6M	0.031	0.033
CANAL	0.009	0.557
PRIMA_PE	-0.051	0.0005
INCREMENTO_ABS	0.052	0.0004
INCREMENTO_REL	0.071	0.00000
PAGOS_EFECTIVO_12M	0.004	0.803
PAGOS_DEBITO_BANCARIO_12M	0.027	0.068
PAGOS_TARJETA_CREDITO_12M	-0.027	0.063
PAGOS_EFECTIVO_6M	0.0002	0.987
PAGOS_DEBITO_BANCARIO_6M	0.030	0.040
PAGOS_TARJETA_CREDITO_6M	-0.031	0.037
PAGOS_EFECTIVO_3M	0.004	0.779
PAGOS_DEBITO_BANCARIO_3M	0.030	0.043
PAGOS_TARJETA_CREDITO_3M	-0.030	0.037

- ❒ INCREMENTO_RELATIVO_2018
- ❒ INCREMENTO_ABSOLUTO_2018
- ❒ DIAS_FIN_CONTRATO
- ❒ SINIESTRALIDAD_12M
- ❒ PRIMA_PE
- ❒ INCREMENTO_ABS
- ❒ INCREMENTO_REL

Algunas variables como MORA_PROM_6M y MORA_MAX_6M se retiraron por estar correlacionadas con MORA_PROM_12M y MORA_MAX_12M, respectivamente, con la diferencia de que las dos últimas no tenían valores perdidos. Para el proceso de eliminación de datos atípicos se decidió retirarlas para no darle un peso excesivo a la mora pero se las reintegrará en el modelamiento.

La distancia de mahalanobis sigue aproximadamente una χ^2 con n grados de libertad, donde n es el número de variables usado, en nuestro caso 13. Se tomó cómo dato atípico a todo aquel que tenga un p-valor menor a 0,001. En función de estas variables se determinó como datos atípicos a 209 de 4677 sujetos, es decir, el 4,5 % de los datos. La tabla siguiente muestra cómo se repartieron estos datos en función de la censura.

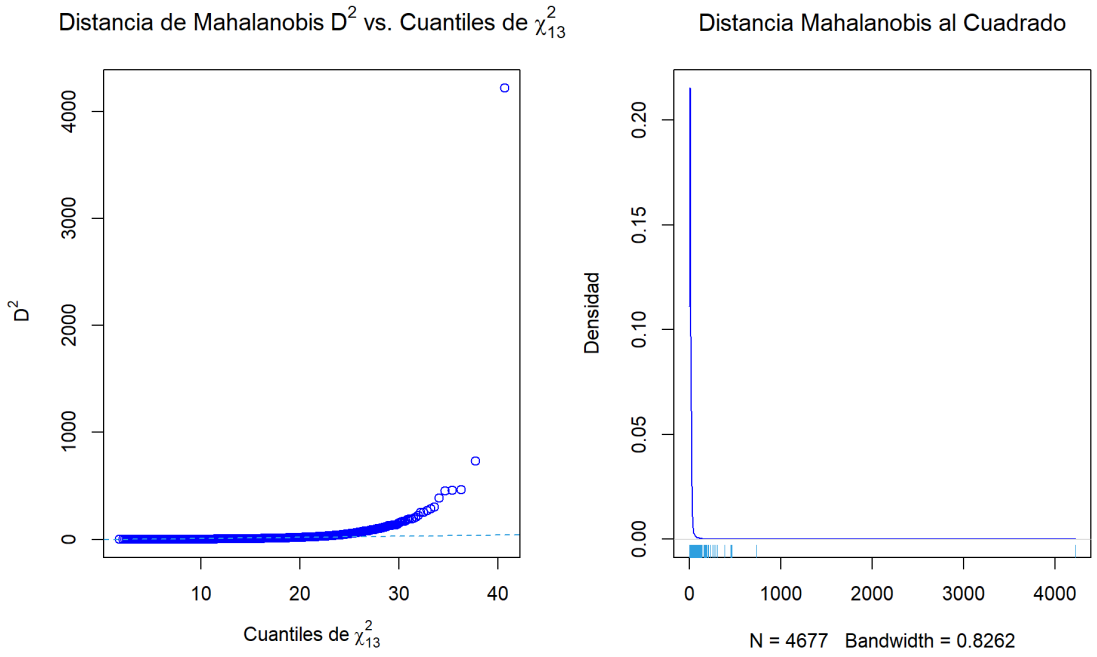
Tabla 4.9: Tabla de contingencia de Censura contra datos atípicos V2

	No Atípico	**Atípico**
Censurado	3319	121
No censurado	1149	88

Mediante una prueba de independencia chi-cuadrado aplicada a la tabla 4.9 podemos determinar si los datos atípicos están distribuidos por igual en los grupo de datos censurados y no censurados o están más concentrados en alguno de ellos. El resultado de la prueba fue el siguiente:

Código de R 4.3: Salida de R prueba chi-cuadrado de la tabla 4.9

```
1         Pearson's Chi-squared test with Yates' continuity correction
-
-  data:  tabla.atipicosV2
-  X-squared = 26.732, df = 1, p-value = 2.337e-07
```



(a) Distancia de Mahalanobis al cuadrado contra cuantiles de la χ^2_{13} correspondiente, en la población V2

(b) función de densidad de la distancia de Mahalanobis para la población V2

Figura 4.6: Distancia de Mahalanobis para la población V2

Cómo el p-valor de la prueba es menor a 0,05 rechazamos la hipótesis nula, entonces podemos decir que los datos atípicos se concentran mayoritariamente en un grupo, el grupo de no censurados para ser más específico.

La figura 4.6a nos muestra cómo se distribuyen los cuantiles de los datos, en caso de tener una distribución perfecta, formarían una línea recta a 45 grados como la que está representada con una línea segmentada. En la figura 4.6b podemos ver la densidad de la distancia de Mahalanobis y donde están mayoritariamente acumulados los datos.

4.5.8 CREACIÓN DE VARIABLES DISCRETAS A PARTIR DE VARIABLES CONTINUAS

La creación de variables discretas, a partir de variables continuas, es una técnica que ha tenido buenos resultados al aplicarse a modelos de riesgos proporcionales. Así lo demuestra el trabajo de Bhandari y Boutros [50] en el que los autores crearon dos modelos de Cox para la predicción de la supervivencia al cáncer de seno. Los análisis de supervivencia asocian el resultado del paciente con una o más variables biológicamente descriptivas. Los

objetivos típicos de tales estudios son evaluar el impacto de un tratamiento o intervención en la supervivencia del paciente a lo largo del plazo, en relación con un grupo de control. Alternativamente, se pueden usar para generar modelos que puedan predecir para cualquier individuo cuál será su riesgo inicial de un evento adverso posterior. Los modelos fueron creados con la misma información y solo diferían en que uno de ellos fue construido con variables continuas únicamente y el otro fue construido solo con variables discretas (Las mismas variables, pero discretizadas). Los resultados obtenidos por el modelo discretizado fueron mejores que los del modelo con las variables continuas y además se pudo comprobar que un modelo mixto, por así llamarlo, tiene mejores resultados que los dos anteriores.

Para la creación de variables discretas se eligió el método de árboles de inferencia condicional, mediante los cual se obtiene la mejor partición de una variable de manera que discrimine los grupos de una variable objetivo. Se uso la función “ctree” incluida en el paquete “partykit”[51] del software R[19]. La metodología que se usa en esta función esta descrita en el capítulo 2 y con mayor detalle en el trabajo Hothorn, Hornik y Zeileis de 2006 [35].

Usando como variable objetivo la censura, podemos obtener las particiones que mejor discriminen los grupos de sujetos: sujetos a los que les ocurrió el evento de interés en un periodo de un año y sujetos a los que no.

Para el presente trabajo se crearon alrededor de 7500 árboles de inferencia condicional con un nivel preespecificado $\alpha = 0,001$, lo que nos asegura que la particiones sean significativas. Debido a la gran cantidad de árboles desarrollados, solo se dará un ejemplo del proceso de discretización para después describir todos los ‘vectores’(llamaremos vectores a las variables discretizadas, para diferenciarlos) que se crearon y cuáles fueron eliminados por estar correlacionados con otro vector creado. Cuatro archivos pdf están accesibles en un repositorio GitHub[12] para que puedan ser revisados por el lector.

Seguidamente, detallaremos el proceso en general. Mediante la función “ctree” se crean los árboles de cada variable e inclusive combinando dos variables. Para este proceso se usó todas las covariables excepto aquellas que fueron eliminadas por tener excesivos datos perdidos y aquellas que carecían de significado. Esta decisión se tomó debido que algunas covariables que no tienen capacidad de discriminar por si solas a los sujetos en estudio, si

[12]link: https://github.com/Qhumir/-rboles-de-decisi-n-Modelo-de-Deserci-n-Clientes

pueden ayudar a particionar otras y de esa forma discriminar de mejor manera los grupos de la variable objetivo. Un ejemplo de código de R usado para este proceso está a continuación:

Código de R 4.4: Ejemplo de código de R para árbol de decisión

```
arbol<- ctree(CENSURA ~ EDAD_TITULAR, data = BaseDatos, controls = ctree_control(
    mincriterion = 0.999))
```

En el código R 4.4 tenemos que $mincriterion = 1 - \alpha$.

Si graficamos el objeto **arbol**[13], resultado del código R 4.4, el producto sería algo como lo que se muestra en la figura 4.7, donde como podemos ver se formaron dos nodos, uno en el que los contratos tenían $EDAD\_TITULAR \leq 70$ y otro en el que $EDAD\_TITULAR > 70$.

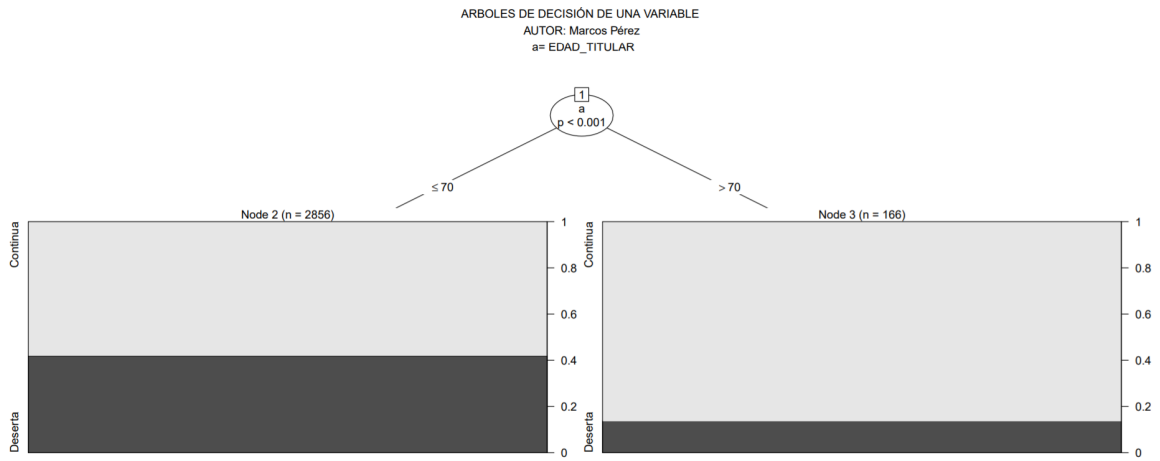


Figura 4.7: Ejemplo de Árbol de Inferencia Condicional

Para discretizar la variable codificamos como 1 al grupo de mayor riesgo y 0 al de menor riesgo, como se muestra en el código R 4.5.

Código de R 4.5: Ejemplo de código de R para discretización

```
vector_5 <- ifelse(  EDAD_TITULAR > 70  , 0, 1)
```

Siguiendo este mismo proceso, y una vez analizados todos los árboles de inferencia condicional, se crearon 76 vectores para la población V1 y 75 para la población V2. No obstante, por volumen de árboles analizados, en algunos casos los vectores se repetían o eran muy parecidos, lo que no nos ayudaba. Debido a lo anterior se eliminó aquellos vectores que

[13] no se tilda debido que en el lenguaje R eso puede causar problemas

estaban correlacionados[14] entre sí, conservando conservando únicamente el de mejores características. Exclusivamente aquellos vectores que entren a los modelos serán descritos a detalle, en cuanto a los demás pueden inferirse de los códigos de R usados para la creación y eliminación de vectores que se presentan a continuación.

Código de R 4.6: Código de R para discretización y eliminación de vectores para población V1

```
################################################################################

############## Vectores V1

attach(BD.V1)

##### Vector 1      V.1V.5
vector_1 <- ifelse(  PRESENTADO_6M_PA > 491.15  , 0, 1)
# nrow(BD)-sum(vector_1);sum(vector_1)   # 113

##### Vector 2      V.1V.22
vector_2 <- ifelse(  RECORTE_3M > 2 , 0, 1)
# nrow(BD)-sum(vector_2);sum(vector_2)   # 266

##### Vector 3      V.1V.24
vector_3 <- ifelse( REEMBOLSOS_CERO_PAGO_9M > 0 , 0, 1)
# nrow(BD)-sum(vector_3);sum(vector_3)   # 345

##### Vector 4      V.1V.25
vector_4 <- ifelse( REEMBOLSOS_CERO_PAGO_12M > 0 , 0, 1)
# nrow(BD)-sum(vector_4);sum(vector_4)   #368

##### Vector 5      V.1V.26
vector_5 <- ifelse(  EDAD_TITULAR > 70 , 0, 1)
# nrow(BD)-sum(vector_5);sum(vector_5)   # 166

##### Vector 6      V.1V.30
vector_6 <- ifelse( NUMERO_AFILIADOS_MENORES <= 0 , 0, 1)
# nrow(BD)-sum(vector_6);sum(vector_6)   # 1998

##### Vector 7      V.1V.31
vector_7 <- ifelse( NUMERO_AFILIADOS_5ANOS_O_MENOS <=  0 , 0, 1)
```

[14]Se usó la correlación de Pearson, que coincide con la de Spearman cuando se comparan dos variables binarias, pero es computacionalmente más eficiente.

```
# nrow(BD)-sum(vector_7);sum(vector_7)   # 2457

##### Vector 8      V.1V.36
vector_8 <- ifelse( TOPE_PLAN > 30000 , 0, 1)
# nrow(BD)-sum(vector_8);sum(vector_8)   # 1658

##### Vector 9      V.1V.56
vector_9 <- ifelse( MOVIMIENTOS_9M <= 1 , 0, 1)
# nrow(BD)-sum(vector_9);sum(vector_9)   # 2366

##### Vector 10      V.1V.57
vector_10 <- ifelse( MOVIMIENTOS_12M <= 1 , 0, 1)
# nrow(BD)-sum(vector_10);sum(vector_10)   # 2255

##### Vector 11      V.1V.75
vector_11 <- ifelse( CANAL <= 0 , 0, 1)
# nrow(BD)-sum(vector_11);sum(vector_11)   # 423

##### Vector 12      V.1V.76
vector_12 <- ifelse( PRIMA_PE > 163.49 , 0, 1)
# nrow(BD)-sum(vector_12);sum(vector_12)   # 266

##### Vector 13      V.1V.78
vector_13 <- ifelse( PAGOS_DEBITO_BANCARIO_12M <= 0.857 , 0, 1)
# nrow(BD)-sum(vector_13);sum(vector_13)   # 1068

##### Vector 14      V.1V.79
vector_14 <- ifelse(  PAGOS_TARJETA_CREDITO_12M > 0 , 0, 1)
# nrow(BD)-sum(vector_14);sum(vector_14)   # 1068

##### Vector 15      V.1V.81
vector_15 <- ifelse(  PAGOS_DEBITO_BANCARIO_6M <= 0.833 , 0, 1)
# nrow(BD)-sum(vector_15);sum(vector_15)   # 1068

##### Vector 16      V.1V.82
vector_16 <- ifelse(  PAGOS_TARJETA_CREDITO_6M > 0 , 0, 1)
# nrow(BD)-sum(vector_16);sum(vector_16)   # 1068

##### Vector 17      V.1V.84
vector_17 <- ifelse( PAGOS_DEBITO_BANCARIO_3M <= 0.667 , 0, 1)
# nrow(BD)-sum(vector_17);sum(vector_17)   # 1068

```

```
##### Vector 18      V.1V.85
vector_18 <- ifelse( PAGOS_TARJETA_CREDITO_3M > 0 , 1, 0)
# nrow(BD)-sum(vector_18);sum(vector_18)   # 1068

##### Vector 19      V.1V.64.2
vector_19 <- ifelse( MORA_MAX_3M <= 10 , 0, 1)
# nrow(BD)-sum(vector_19);sum(vector_19)   # 2617

##### Vector 20      V.2V.60.5
vector_20 <- ifelse( MORA_PROM_6M > 4.6 , 1, 0)
# nrow(BD)-sum(vector_20);sum(vector_20)   # 364

##### Vector 24      V.1V.59.3
vector_24 <- ifelse( MORA_PROM_12M <= 0.75 , 0, 1)
# nrow(BD)-sum(vector_24);sum(vector_24)   # 2282

##### Vector 25      V.2V.3295.3
vector_25 <- ifelse( MORA_MAX_6M <= 2 & MORA_PROM_3M <= 5 , 0, 1)
# nrow(BD)-sum(vector_25);sum(vector_25)   # 2245

##### Vector 26      V.2V.1910.4
vector_26 <- ifelse( EDAD_TITULAR <= 65 & PAGOS_DEBITO_BANCARIO_12M > .857    , 1,
    0)
# nrow(BD)-sum(vector_26);sum(vector_26)   # 1831

##### Vector 29      V.2V.3167.3
vector_29 <- ifelse( MOVIMIENTOS_9M <= 0 & MORA_MAX_12M <= 10 , 0, 1)
# nrow(BD)-sum(vector_29);sum(vector_29)   # 1383

##### Vector 30      V.2V.3332.11 # Se cambió por mejor Opción
vector_30 <- ifelse( MORA_PROM_3M > 5 & PAGOS_DEBITO_BANCARIO_12M > 0 , 1, 0)
# nrow(BD)-sum(vector_30);sum(vector_30)   # 173 + 142

##### Vector 31      V.2V.3333.10
vector_31 <- ifelse( MORA_PROM_3M > 11.333 & PAGOS_TARJETA_CREDITO_12M <= 0 , 1,
    0)
# nrow(BD)-sum(vector_31);sum(vector_31)   # 173

##### Vector 34      V.2V.3261.3 # Se cambió por mejor Opción
vector_34 <- ifelse( MORA_MAX_12M <= 10 , 0, 1)
# nrow(BD)-sum(vector_34);sum(vector_34)   # 2517

```

```
##### Vector 35      V.2V.3308.3 # Se cambió por mejor Opción
vector_35 <- ifelse( MORA_MAX_6M <= 10 , 0, 1)
# nrow(BD)-sum(vector_35);sum(vector_35)   # 2289

##### Vector 36      V.2V.3098.5
vector_36 <- ifelse( CAMBIO_PLAN_12M > 0 & PAGOS_DEBITO_BANCARIO_12M > .857 , 1,
    0)
# nrow(BD)-sum(vector_36);sum(vector_36)   # 940

##### Vector 39      V.2V.1907.4
vector_39 <- ifelse( EDAD_TITULAR <= 70 & CANAL > 0 , 1, 0)
# nrow(BD)-sum(vector_39);sum(vector_39)   # 2487

##### Vector 41      V.2V.3516.4
vector_41 <- ifelse( CANAL > 0 & PRIMA_PE <= 163.49 , 1, 0)
# nrow(BD)-sum(vector_41);sum(vector_41)   # 2410

##### Vector 42      V.2V.3329.5
vector_42 <- ifelse( MORA_PROM_3M <= .75 & CANAL > 0 , 0, 1)
# nrow(BD)-sum(vector_42);sum(vector_42)   # 2029

##### Vector 43      V.1V.62.5
vector_43 <- ifelse( MORA_MAX_6M > 39 , 1, 0)
# nrow(BD)-sum(vector_43);sum(vector_43)   # 156

##### Vector 44      V.2V.2237.5
vector_44 <- ifelse( INDIVIDUAL_FAMILIAR > 1 & PAGOS_DEBITO_BANCARIO_12M > .857 ,
     1, 0)
# nrow(BD)-sum(vector_44);sum(vector_44)   # 687

##### Vector 49      V.2V.1742.3
vector_49 <- ifelse( REEMBOLSOS_CERO_PAGO_9M <= 0 & EDAD_TITULAR <= 70 , 1, 0)
# nrow(BD)-sum(vector_49);sum(vector_49)   # 2530

##### Vector 50      V.2V.1791.3
vector_50 <- ifelse( REEMBOLSOS_CERO_PAGO_9M <= 0 & PRIMA_PE <= 163.49 , 1, 0)
# nrow(BD)-sum(vector_50);sum(vector_50)   # 2446

##### Vector 52      V.2V.3285.6
vector_52 <- ifelse( MORA_PROM_6M <= 4.6 & PRIMA_PE > 160.82 , 0, 1)
# nrow(BD)-sum(vector_52);sum(vector_52)   # 255

```

```
##### Vector 53      V.2V.1801.3
vector_53 <- ifelse( REEMBOLSOS_CERO_PAGO_12M <= 0 & EDAD_TITULAR <= 70 , 1, 0)
# nrow(BD)-sum(vector_53);sum(vector_53)   # 2507

##### Vector 54      V.2V.3212.5
vector_54 <- ifelse( MOVIMIENTOS_12M > 1 & PAGOS_DEBITO_BANCARIO_12M > .875 , 1,
    0)
# nrow(BD)-sum(vector_54);sum(vector_54)   # 471

##### Vector 59      V.2V.2432.4
vector_59 <- ifelse( TOPE_PLAN <= 30000 & CANAL > 0 , 1, 0)
# nrow(BD)-sum(vector_59);sum(vector_59)   # 1198

##### Vector 60      V.2V.1745.4
vector_60 <- ifelse( REEMBOLSOS_CERO_PAGO_9M <= 0 & NUMERO_AFILIADOS_MENORES > 0
    , 1, 0)
# nrow(BD)-sum(vector_60);sum(vector_60)   # 835

##### Vector 61      V.2V.1804.4
vector_61 <- ifelse( REEMBOLSOS_CERO_PAGO_12M <= 0 & NUMERO_AFILIADOS_MENORES > 0
     , 1, 0)
# nrow(BD)-sum(vector_61);sum(vector_61)   # 818

##### Vector 62      V.2V.194.4
vector_62 <- ifelse( PRESENTADO_3M_PA <= 0 & NUMERO_AFILIADOS_MENORES > 0 , 1, 0)
# nrow(BD)-sum(vector_62);sum(vector_62)   # 844

##### Vector 64      V.2V.2037.4
vector_64 <- ifelse( NUMERO_AFILIADOS_MENORES > 0 & TOPE_PLAN <= 30000 , 1, 0)
# nrow(BD)-sum(vector_64);sum(vector_64)   # 444

##### Vector 65      V.2V.1748.4
vector_65 <- ifelse( REEMBOLSOS_CERO_PAGO_9M <= 0 & INDIVIDUAL_FAMILIAR > 1 , 1,
    0)
# nrow(BD)-sum(vector_65);sum(vector_65)   # 968

##### Vector 66      V.2V.1805.4
vector_66 <- ifelse( REEMBOLSOS_CERO_PAGO_12M <= 0 &
    NUMERO_AFILIADOS_5ANOS_O_MENOS > 0 , 1, 0)
# nrow(BD)-sum(vector_66);sum(vector_66)   # 408

##### Vector 67      V.2V.1807.4
```

```
vector_67 <- ifelse( REEMBOLSOS_CERO_PAGO_12M <= 0 & INDIVIDUAL_FAMILIAR > 1 , 1,
    0)
# nrow(BD)-sum(vector_67);sum(vector_67)   # 960

##### Vector 68      V.2V.1746.4
vector_68 <- ifelse( REEMBOLSOS_CERO_PAGO_9M <= 0 &
   NUMERO_AFILIADOS_5ANOS_O_MENOS > 0 , 1, 0)
# nrow(BD)-sum(vector_68);sum(vector_68)   # 423

##### Vector 69      V.2V.3182.5
vector_69 <- ifelse( MOVIMIENTOS_9M > 1 & CANAL > 0 , 1, 0)
# nrow(BD)-sum(vector_69);sum(vector_69)   # 584

##### Vector 71      V.2V.2091.4
vector_71 <- ifelse( NUMERO_AFILIADOS_5ANOS_O_MENOS > 0 & TOPE_PLAN <= 30000 , 1,
    0)
# nrow(BD)-sum(vector_71);sum(vector_71)   # 223

##### Vector 72      V.2V.2078.5
vector_72 <- ifelse( NUMERO_AFILIADOS_MENORES > 1 & PAGOS_DEBITO_BANCARIO_12M >
   .857 , 1, 0)
# nrow(BD)-sum(vector_72);sum(vector_72)   # 235

##### Vector 76      V.2V.3332.6
vector_76 <- ifelse( MORA_PROM_3M <= 5 & MORA_PROM_3M > .667 &
   PAGOS_DEBITO_BANCARIO_12M > .857 , 1, 0)
# nrow(BD)-sum(vector_76);sum(vector_76)   # 192

cen <- CENSURA

detach(BD.V1)


################################################################################

############## Eliminación de Vectores Correlacionados V1

a <- NULL
b <- NULL
k=0
for(i in 1:99){
    if( !(exists(paste0("vector_",i))) ){next}
```

```
    k=k+1
    if(i %n% b){next}
    for(j in (i+1):100){
        if( !(exists(paste0("vector_",j))) ){next}
        if(abs(cor( get(paste0("vector_",i)) , get(paste0("vector_",j)) ))>=.80){
        if(abs(cor( get(paste0("vector_",i)), cen)) >= abs(cor( get(paste0("
            vector_",j)) , cen))){kk <- j
        }else{kk <- i}
        b <- c(b,kk)
        }
    }
}
vectores.finales.V1 <- unique(b[order(b)])
```

Código de R 4.7: Código de R para discretización y eliminación de vectores para población V2

```
################################################################################

############## Vectores V2

attach(BD.V2)

##### Vector 1      V.1V.21
vector_1 <- ifelse(  RECORTE_6M <= 116.5  , 0, 1)
# nrow(BD)-sum(vector_1);sum(vector_1)   # 4307

##### Vector 2      V.1V.22
vector_2 <- ifelse(  RECORTE_3M <= 61.61  , 0, 1)
# nrow(BD)-sum(vector_2);sum(vector_2)   # 4365

##### Vector 3      V.1V.27
vector_3 <- ifelse(  EDAD_TITULAR > 49  , 0, 1)
# nrow(BD)-sum(vector_3);sum(vector_3)   # 1264

##### Vector 4      V.1V.30
vector_4 <- ifelse(  NUMERO_AFILIADOS_MENORES <= 0  , 0, 1)
# nrow(BD)-sum(vector_4);sum(vector_4)   # 2726

##### Vector 5      V.1V.31
vector_5 <- ifelse(  NUMERO_AFILIADOS_5ANOS_O_MENOS <= 0  , 0, 1)
# nrow(BD)-sum(vector_5);sum(vector_5)   # 3626

```

```
##### Vector 6      V.1V.32
vector_6 <- ifelse( NUMERO_AFILIADOS_2ANOS_O_MENOS <= 0 , 0, 1)
# nrow(BD)-sum(vector_6);sum(vector_6)   # 4130

##### Vector 7      V.1V.35
vector_7 <- ifelse( TIEMPO_AFILIACION > 1484 , 0, 1)
# nrow(BD)-sum(vector_7);sum(vector_7)   # 1191

##### Vector 8      V.1V.36
vector_8 <- ifelse(  PRODUCTO_MH > 0 , 0, 1)
# nrow(BD)-sum(vector_8);sum(vector_8)   # 2414

##### Vector 9      V.1V.38
vector_9 <- ifelse(  TITULAR_SIN_BENEFICIOS <= 0 , 0, 1)
# nrow(BD)-sum(vector_9);sum(vector_9)   # 3510

##### Vector 10      V.1V.59
vector_10 <- ifelse(  MORA_PROM_12M <= 5.083 , 0, 1)
# nrow(BD)-sum(vector_10);sum(vector_10)   # 4049

##### Vector 11      V.1V.64
vector_11 <- ifelse(  MORA_MAX_3M <= 14 , 0, 1)
# nrow(BD)-sum(vector_11);sum(vector_11)   # 4095

##### Vector 12      V.1V.65
vector_12 <- ifelse( INCREMENTO_RELATIVO_2018 <= .156 , 0, 1)
# nrow(BD)-sum(vector_12);sum(vector_12)   # 3396

##### Vector 13      V.1V.67
vector_13 <- ifelse(  DIAS_FIN_CONTRATO <= 311 , 0, 1)
# nrow(BD)-sum(vector_13);sum(vector_13)   # 4234

##### Vector 14      V.1V.74
vector_14 <- ifelse(   SINIESTRALIDAD_12M <= .513 , 0, 1)
# nrow(BD)-sum(vector_14);sum(vector_14)   # 3876

##### Vector 15      V.1V.75
vector_15 <- ifelse(  SINIESTRALIDAD_6M <= 0.763 , 0, 1)
# nrow(BD)-sum(vector_15);sum(vector_15)   # 4342

##### Vector 16      V.1V.77
vector_16 <- ifelse(  PRIMA_PE >  86.183 , 0, 1)
```

```
# nrow(BD)-sum(vector_16);sum(vector_16)   # 1503

##### Vector 17      V.1V.79
vector_17 <- ifelse( INCREMENTO_REL <= 0.156 , 0, 1)
# nrow(BD)-sum(vector_17);sum(vector_17)   # 3402

##### Vector 18      V.2V.3489.4
vector_18 <- ifelse( MORA_MAX_6M <= 14 & SINIESTRALIDAD_12M <= .516 , 0, 1)
# nrow(BD)-sum(vector_18);sum(vector_18)   # 3440

##### Vector 19      V.2V.3531.3
vector_19 <- ifelse( MORA_MAX_3M <= 14 & DIAS_FIN_CONTRATO <= 318 , 0, 1)
# nrow(BD)-sum(vector_19);sum(vector_19)   # 3916

##### Vector 20      V.2V.3538.3
vector_20 <- ifelse( MORA_MAX_3M <= 14 & SINIESTRALIDAD_12M <= .516 , 0, 1)
# nrow(BD)-sum(vector_20);sum(vector_20)   # 3564

##### Vector 21      V.2V.3436.3
vector_21 <- ifelse( MORA_MAX_12M <= 31 & SINIESTRALIDAD_12M <= .516 , 0, 1)
# nrow(BD)-sum(vector_21);sum(vector_21)   # 3543

##### Vector 22      V.2V.3426.3
vector_22 <- ifelse( MORA_MAX_12M <= 31 & MORA_MAX_3M <= 14 , 0, 1)
# nrow(BD)-sum(vector_22);sum(vector_22)   # 3943

##### Vector 23      V.2V.2264.3
vector_23 <- ifelse( NUMERO_AFILIADOS_2ANOS_O_MENOS <= 0 & MORA_MAX_3M <= 14 , 0,
    1)
# nrow(BD)-sum(vector_23);sum(vector_23)   # 3787

##### Vector 24      V.2V.2263.3
vector_24 <- ifelse( NUMERO_AFILIADOS_2ANOS_O_MENOS <= 0 & MORA_PROM_3M <= 5.3334
    , 0, 1)
# nrow(BD)-sum(vector_24);sum(vector_24)   # 3778

##### Vector 25      V.2V.2262.4
vector_25 <- ifelse( NUMERO_AFILIADOS_2ANOS_O_MENOS <= 0 & MORA_MAX_6M <= 6 , 0,
    1)
# nrow(BD)-sum(vector_25);sum(vector_25)   # 3448

##### Vector 26      V.2V.2260.3
```

```
vector_26 <- ifelse( NUMERO_AFILIADOS_2ANOS_O_MENOS <= 0 & MORA_MAX_12M <= 31 ,
    0, 1)
# nrow(BD)-sum(vector_26);sum(vector_26)   # 3776

##### Vector 27      V.2V.3494.4
vector_27 <- ifelse( MORA_MAX_6M <= 6 & INCREMENTO_REL <= .156 , 0, 1)
# nrow(BD)-sum(vector_27);sum(vector_27)   # 2853

##### Vector 28      V.2V.1593.3
vector_28 <- ifelse( RECORTE_6M <= 120 & MORA_MAX_3M <= 14 , 0, 1)
# nrow(BD)-sum(vector_28);sum(vector_28)   # 3949

##### Vector 29      V.2V.1658.3
vector_29 <- ifelse( RECORTE_3M <= 61.61 & MORA_PROM_3M <= 5.3334 , 0, 1)
# nrow(BD)-sum(vector_29);sum(vector_29)   # 3987

##### Vector 30      V.2V.1945.3
vector_30 <- ifelse( EDAD_TITULAR <= 49 & TIEMPO_AFILIACION <= 1484 , 1, 0)
# nrow(BD)-sum(vector_30);sum(vector_30)   # 2588

##### Vector 31      V.2V.1592.3
vector_31 <- ifelse( RECORTE_6M <= 116.5 & MORA_PROM_3M <= 5.3334 , 0, 1)
# nrow(BD)-sum(vector_31);sum(vector_31)   # 3935

##### Vector 32      V.2V.3490.4
vector_32 <- ifelse( MORA_MAX_6M <= 14 & SINIESTRALIDAD_6M <= .095 , 0, 1)
# nrow(BD)-sum(vector_32);sum(vector_32)   # 2948

##### Vector 33      V.2V.3464.3
vector_33 <- ifelse( MORA_PROM_6M <= 5.8334 & SINIESTRALIDAD_6M <= .855 , 0, 1)
# nrow(BD)-sum(vector_33);sum(vector_33)   # 4000

##### Vector 34      V.2V.2423.4
vector_34 <- ifelse( TIEMPO_AFILIACION > 1484 & MORA_PROM_6M <= 5.8334 , 0, 1)
# nrow(BD)-sum(vector_34);sum(vector_34)   # 1126

##### Vector 35      V.2V.60.3
vector_35 <- ifelse( PRESENTADO <= 1274.67 & MORA_PROM_6M <= 5.8334 , 0, 1)
# nrow(BD)-sum(vector_35);sum(vector_35)   # 4003

##### Vector 36      V.2V.875.3
vector_36 <- ifelse( PRESENTADO_3M <= 1274.67 & MORA_PROM_6M <= 5.8334 , 0, 1)
```

```
# nrow(BD)-sum(vector_36);sum(vector_36)   # 4003

##### Vector 37      V.2V.1589.3
vector_37 <- ifelse( RECORTE_6M <= 120 & MORA_MAX_12M <= 31 , 0, 1)
# nrow(BD)-sum(vector_37);sum(vector_37)   # 3944

##### Vector 38      V.1V.60.3
vector_38 <- ifelse( MORA_MAX_12M <= 6 , 0, 1)
# nrow(BD)-sum(vector_38);sum(vector_38)   # 3469

##### Vector 39      V.2V.2205.3
vector_39 <- ifelse( NUMERO_AFILIADOS_5ANOS_O_MENOS <= 0 & MORA_PROM_6M <= 5.8334
    , 0, 1)
# nrow(BD)-sum(vector_39);sum(vector_39)   # 3341

##### Vector 40      V.2V.2578.3
vector_40 <- ifelse( TITULAR_SIN_BENEFICIOS <= 0 & MORA_PROM_3M <= 5.3334 , 0, 1)
# nrow(BD)-sum(vector_40);sum(vector_40)   # 3239

##### Vector 41      V.2V.1984.4
vector_41 <- ifelse( EDAD_TITULAR <= 49 & SINIESTRALIDAD_12M > .116 , 1, 0)
# nrow(BD)-sum(vector_41);sum(vector_41)   # 1505

##### Vector 42      V.2V.3440.3
vector_42 <- ifelse( MORA_MAX_12M <= 31 & INCREMENTO_ABS <= 16.73 , 0, 1)
# nrow(BD)-sum(vector_42);sum(vector_42)   # 2986

##### Vector 43      V.2V.232.4
vector_43 <- ifelse( PRESENTADO_3M_PA <= 39.89 & MORA_MAX_6M <= 6 , 0, 1)
# nrow(BD)-sum(vector_43);sum(vector_43)   # 3044

##### Vector 44      V.2V.2149.4
vector_44 <- ifelse( NUMERO_AFILIADOS_MENORES <= 0 & MORA_MAX_6M <= 6 , 0, 1)
# nrow(BD)-sum(vector_44);sum(vector_44)   # 2336

##### Vector 45      V.2V.2427.4
vector_45 <- ifelse( TIEMPO_AFILIACION <= 1484 & INCREMENTO_RELATIVO_2018 > .19 ,
    1, 0)
# nrow(BD)-sum(vector_45);sum(vector_45)   # 675

##### Vector 46      V.2V.2151.3
vector_46 <- ifelse( NUMERO_AFILIADOS_MENORES <= 0 & MORA_MAX_3M <= 14 , 0, 1)
```

```
# nrow(BD)-sum(vector_46);sum(vector_46)   # 2540

##### Vector 47      V.2V.2441.4
vector_47 <- ifelse( TIEMPO_AFILIACION <= 1484 & INCREMENTO_REL > .19 , 1, 0)
# nrow(BD)-sum(vector_47);sum(vector_47)   # 666

##### Vector 48      V.2V.3554.3
vector_48 <- ifelse( INCREMENTO_RELATIVO_2018 <= .156 & DIAS_FIN_CONTRATO <= 318
    , 0, 1)
# nrow(BD)-sum(vector_48);sum(vector_48)   # 3229

##### Vector 49      V.2V.3609.3
vector_49 <- ifelse( DIAS_FIN_CONTRATO <= 318 & INCREMENTO_REL <= .156 , 0, 1)
# nrow(BD)-sum(vector_49);sum(vector_49)   # 3234

##### Vector 50      V.2V.1988.5
vector_50 <- ifelse( EDAD_TITULAR <= 39 & INCREMENTO_ABS > 12.97 , 1, 0)
# nrow(BD)-sum(vector_50);sum(vector_50)   # 487

##### Vector 51      V.2V.2589.3
vector_51 <- ifelse( TITULAR_SIN_BENEFICIOS <= 0 & SINIESTRALIDAD_12M <= .513 ,
    0, 1)
# nrow(BD)-sum(vector_51);sum(vector_51)   # 3078

##### Vector 52      V.2V.2478.4
vector_52 <- ifelse( PRODUCTO_MH > 0 & MORA_MAX_3M <= 14 , 0, 1)
# nrow(BD)-sum(vector_52);sum(vector_52)   # 2230

##### Vector 53      V.2V.2218.3
vector_53 <- ifelse( NUMERO_AFILIADOS_5ANOS_O_MENOS <= 0 & SINIESTRALIDAD_12M <=
    .513 , 0, 1)
# nrow(BD)-sum(vector_53);sum(vector_53)   # 3199

##### Vector 54      V.2V.3478.7
vector_54 <- ifelse( MORA_MAX_6M > 23 & MORA_PROM_3M > 11.667 , 1, 0)
# nrow(BD)-sum(vector_54);sum(vector_54)   # 180

##### Vector 55      V.2V.2733.5
vector_55 <- ifelse( PRESENTADO_CPC_12M <= 328.62 & SINIESTRALIDAD_12M <= .513 &
    SINIESTRALIDAD_12M > .116 , 1, 0)
# nrow(BD)-sum(vector_55);sum(vector_55)   # 538

```

```
##### Vector 56      V.2V.2580.3
vector_56 <- ifelse( TITULAR_SIN_BENEFICIOS <= 0 & INCREMENTO_RELATIVO_2018 <=
    .156 , 0, 1)
# nrow(BD)-sum(vector_56);sum(vector_56)   # 2695

##### Vector 57      V.2V.2594.3
vector_57 <- ifelse( TITULAR_SIN_BENEFICIOS <= 0 & INCREMENTO_REL <= .156 , 0, 1)
# nrow(BD)-sum(vector_57);sum(vector_57)   # 2701

##### Vector 58      V.2V.3726.4
vector_58 <- ifelse( SINIESTRALIDAD_12M <= .513 & PRIMA_PE > 84.73 , 0, 1)
# nrow(BD)-sum(vector_58);sum(vector_58)   # 1358

##### Vector 60      V.2V.2267.3
vector_60 <- ifelse( NUMERO_AFILIADOS_2ANOS_O_MENOS <= 0 & DIAS_FIN_CONTRATO <=
    304 , 0, 1)
# nrow(BD)-sum(vector_60);sum(vector_60)   # 3892

##### Vector 61      V.2V.3727.5
vector_61 <- ifelse( SINIESTRALIDAD_12M > .513 & INCREMENTO_ABS > 6.72 , 1, 0)
# nrow(BD)-sum(vector_61);sum(vector_61)   # 382

##### Vector 62      V.2V.3439.3   #Se cambió por mejor opción
vector_62 <- ifelse( MORA_MAX_12M <= 31 , 0, 1)
# nrow(BD)-sum(vector_62);sum(vector_62)   # 3874

##### Vector 63      V.2V.2733.6
vector_63 <- ifelse( PRESENTADO_CPC_12M > 328.62 & SINIESTRALIDAD_12M <= .513 ,
    0, 1)
# nrow(BD)-sum(vector_63);sum(vector_63)   # 1262

##### Vector 64      V.2V.3605.3
vector_64 <- ifelse( DIAS_FIN_CONTRATO <= 311 & SINIESTRALIDAD_6M <= .763 , 0, 1)
# nrow(BD)-sum(vector_64);sum(vector_64)   # 4116

##### Vector 65      V.2V.3517.4
vector_65 <- ifelse( MORA_PROM_3M <= 5.3334 & PRIMA_PE > 104.675 , 0, 1)
# nrow(BD)-sum(vector_65);sum(vector_65)   # 935

##### Vector 66      V.2V.1662.3
vector_66 <- ifelse( RECORTE_3M <= 61.61 & DIAS_FIN_CONTRATO <= 311 , 0, 1)
# nrow(BD)-sum(vector_66);sum(vector_66)   # 4132
```

```
-
- ##### Vector 67      V.1V.61.5
- vector_67 <- ifelse( MORA_PROM_6M > 15.167, 1, 0)
- # nrow(BD)-sum(vector_67);sum(vector_67)   # 106
270
- ##### Vector 68      V.2V.2275.3
- vector_68 <- ifelse( NUMERO_AFILIADOS_2ANOS_O_MENOS <= 0 & SINIESTRALIDAD_6M <=
    .763 , 0, 1)
- # nrow(BD)-sum(vector_68);sum(vector_68)   # 4023
-
275 ##### Vector 69      V.2V.2779.8
- vector_69 <- ifelse( PAGADO_NETO_CPC_12M > 608.44 & SINIESTRALIDAD_12M <= .513 ,
    0, 1)
- # nrow(BD)-sum(vector_69);sum(vector_69)   # 530
-
- ##### Vector 70      V.2V.3564.4
280 vector_70 <- ifelse( INCREMENTO_RELATIVO_2018 <= .156 & PRIMA_PE > 115.79 , 0, 1)
- # nrow(BD)-sum(vector_70);sum(vector_70)   # 544
-
- ##### Vector 71      V.2V.2123.4
- vector_71 <- ifelse( NUMERO_AFILIADOS_MENORES > 0 & PRODUCTO_MH <= 0 , 1, 0)
285 # nrow(BD)-sum(vector_71);sum(vector_71)   # 673
-
- ##### Vector 72      V.2V.2166.5
- vector_72 <- ifelse( NUMERO_AFILIADOS_MENORES > 1 & INCREMENTO_REL > .156 , 1, 0)
- # nrow(BD)-sum(vector_72);sum(vector_72)   # 158
290
- ##### Vector 73      V.2V.2638.6
- vector_73 <- ifelse( PRESENTADO_CRONICO_12M > 619.28 & SINIESTRALIDAD_12M <= .513
     , 0, 1)
- # nrow(BD)-sum(vector_73);sum(vector_73)   # 350
-
295 ##### Vector 74      V.2V.3541.4
- vector_74 <- ifelse( MORA_MAX_3M <= 14 & PRIMA_PE > 136.055 , 0, 1)
- # nrow(BD)-sum(vector_74);sum(vector_74)   # 469
-
- ##### Vector 75      V.2V.1977.4
300 vector_75 <- ifelse( EDAD_TITULAR <= 49 & DIAS_FIN_CONTRATO > 318 , 1, 0)
- # nrow(BD)-sum(vector_75);sum(vector_75)   # 166
-
- cen <- CENSURA
-
```

```r
detach(BD.V2)



################################################################################

############## Eliminación de Vectores Correlacionados V2

a <- NULL
b <- NULL
k=0
for(i in 1:99){
    if( !(exists(paste0("vector_",i))) ){next}
    k=k+1
    if(i %n% b){next}
    for(j in (i+1):100){
        if( !(exists(paste0("vector_",j))) ){next}
        if(abs(cor( get(paste0("vector_",i)) , get(paste0("vector_",j)) ))>=.70){
        if(abs(cor( get(paste0("vector_",i)), cen)) >= abs(cor( get(paste0("
            vector_",j)) , cen))){kk <- j
        }else{kk <- i}
        b <- c(b,kk)
        }
    }
}
vectores.finales.V2 <- unique(b[order(b)])
```

4.5.9 ESTIMACIÓN DEL MODELO DE COX

La estimación del modelo de riesgo proporcionales se realizó mediante la función “coxph” incluida en el paquete “survival”[18] del software R[19]. Tras descartar variables -por cumplir o no los supuestos del modelo- llegamos a dos modelos finales, uno para cada población. Considerando que la **hipótesis de riesgo proporcionales** es un eje fundamental del modelo Cox, los resultados de esta prueba se los presentará en un apartado diferente y los demás supuestos por separado.

4.5.9.1 Estimación del modelo de Cox para la población V1

La estimación del modelo de Cox para la población V1 se inició con las variables que en el proceso de depuración se consideró que tenían buena capacidad de discriminación junto con todos los vectores construidos que no se eliminó. Las covariables que se tomaron en cuenta fueron las siguientes:

- ❒ PRESENTADO_6M
- ❒ PAGADO_NETO_6M
- ❒ PAGADO_NETO_12M
- ❒ EDAD_TITULAR
- ❒ NUMERO_AFILIADOS
- ❒ INDIVIDUAL_FAMILIAR
- ❒ TOPE_PLAN
- ❒ CAMBIO_PLAN_9M
- ❒ CAMBIO_PLAN_12M
- ❒ MOVIMIENTOS_9M
- ❒ MOVIMIENTOS_12M
- ❒ USO_6M
- ❒ TIEMPO_SIN_USO
- ❒ CANAL
- ❒ PRIMA_PE
- ❒ PAGOS_DEBITO_BANCARIO_12M
- ❒ PAGOS_TARJETA_CREDITO_12M
- ❒ PAGOS_DEBITO_BANCARIO_6M
- ❒ PAGOS_TARJETA_CREDITO_6M
- ❒ PAGOS_DEBITO_BANCARIO_3M
- ❒ PAGOS_TARJETA_CREDITO_3M
- ❒ vector_1
- ❒ vector_2
- ❒ vector_5
- ❒ vector_26
- ❒ vector_29

- ❒ vector_30
- ❒ vector_31
- ❒ vector_36
- ❒ vector_39
- ❒ vector_42
- ❒ vector_44
- ❒ vector_49
- ❒ vector_52
- ❒ vector_54
- ❒ vector_59
- ❒ vector_60
- ❒ vector_64
- ❒ vector_65
- ❒ vector_66
- ❒ vector_69
- ❒ vector_71
- ❒ vector_72
- ❒ vector_76

Recordemos que, como se dijo antes, solo se describirá a detalle los vectores que ingresen al modelo.

En base a las variables nombradas y a un arduo proceso de modelización se tuvo como resultado un modelo de riesgos proporcionales que cumple con todos los requisitos y supuestos. A continuación, se presenta el código de R y luego el resultado final:

Código de R 4.8: Código Modelo de Cox población V1

```
#############################################################################

############## Modelo de Cox 1

Modelo_Cox1 <- coxph(formula = Surv(TIEMPO, CENSURA) ~ CANAL + EDAD_TITULAR +
    vector_1 + vector_2 + vector_5 + vector_36 + vector_44 + vector_71,
    data = Base_Cox, y = TRUE, x = TRUE)

cox.zph(Modelo_Cox1)
```

En la tabla 4.10 podemos encontrar un resumen del modelo V1, entre los datos que podemos ver están los coeficientes de cada una de la covariables que ingresaron al modelo final junto con sus errores estándar y su significancia, los estadísticos para el contraste de hipótesis, el índice de concordancia (Concordance) y el R^2.

Los estadísticos de las prueba de contraste de hipótesis de Wald, razón de verosimilitud

(LR) y Score tienen un p-valor por debajo de 0,01. En resumen, todas las pruebas de contraste nos dicen que el modelo está bien ajustado.

Tabla 4.10: Resumen del Modelos de Cox para la población V1

	Dependent variable:
	TIEMPO
CANAL	0.424***
	(0.100)
EDAD_TITULAR	−0.005*
	(0.003)
vector_1	0.429**
	(0.208)
vector_2	0.490***
	(0.129)
vector_5	0.924***
	(0.237)
vector_36	0.321***
	(0.063)
vector_44	0.280***
	(0.070)
vector_71	0.408***
	(0.097)
Observations	3,022
Concordance	0.609 (0.008)
R^2	0.068
Max. Possible R^2	0.998
Log Likelihood	−9,390.266
Wald Test	189.420*** (df = 8)
LR Test	211.689*** (df = 8)
Score (Logrank) Test	200.599*** (df = 8)
Note:	*p<0.1; **p<0.05; ***p<0.01

El modelo final tiene un índice de concordancia de 0,61 lo que es bueno, sin embargo, también tiene un R^2 poco prometedor (0,068). Basado en esto podemos decir que el modelo tiene poca capacidad predictiva, pero si nos puede dar una idea de las variables que influyen mayoritariamente en la deserción de clientes.

Antes de hacer un análisis del modelo, se explicará que es lo que mide cada covariable que ingresó al modelo:

1. **CANAL**. Variable binaria. Toma el valor de 1 si el canal el contrato fue suscrito fue a

través del canal directo, es decir, fue suscrito por un ejecutivo de ventas de Humana S.A., y 0 en caso de haber sido suscrito por otros canales (Bróker o página web).

2. **EDAD_TITULAR**. Como su nombre lo indica nos dice la edad del titular del contrato.

3. **vector_1**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando valor presentado (ver anexo 7.3)por los afiliados del contrato en la modalidad de pago al afiliado (ver anexo 7.3) en los últimos 6 meses es superior a $491,15 y 1 (mayor riesgo de deserción) en caso contrario.

4. **vector_2**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando la suma total de los recortes en los últimos 3 meses es mayor a $2,00 y 1 (mayor riesgo de deserción) en caso contrario.

5. **vector_5**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando la edad del titular es mayor a 70 y 1 (mayor riesgo de deserción) en caso contrario.

6. **vector_36**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando ha habido cambios de plan en los últimos 12 meses y más del 85,7 % de las primas del contrato han sido pagadas por débito bancario en los últimos 12 meses. Toma el valor de 0 (menor riesgo de deserción) en caso contrario.

7. **vector_44**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando se trata de un contrato familiar -es decir, con más de una persona en él- y más del 85,7 % de las primas del contrato han sido pagadas por débito bancario en los últimos 12 meses. Toma el valor de 0 (menor riesgo de deserción) en caso contrario.

8. **vector_71**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando hay al menos un afiliado menor de 5 años y el tope de cobertura es menor o igual a $30.000,00. Toma el valor de 0 (menor riesgo de deserción) en caso contrario.

También podemos ver la figura 4.8 que nos muestra el intervalo de confianza de la exponencial de los coeficientes del modelo ($\exp(\hat{\beta})$), poniéndolos en perspectiva unos con otros. Nos será útil para analizar cada covariable que ingresó al modelo.

La variable que más peso tiene en el modelo es el vector_5. Esto nos dice que la Edad del titular influye mucho cuando el cliente no ha superado el primer año, a una edad mayor a 70 años le corresponde una menor probabilidad de deserción. Esto puede deberse

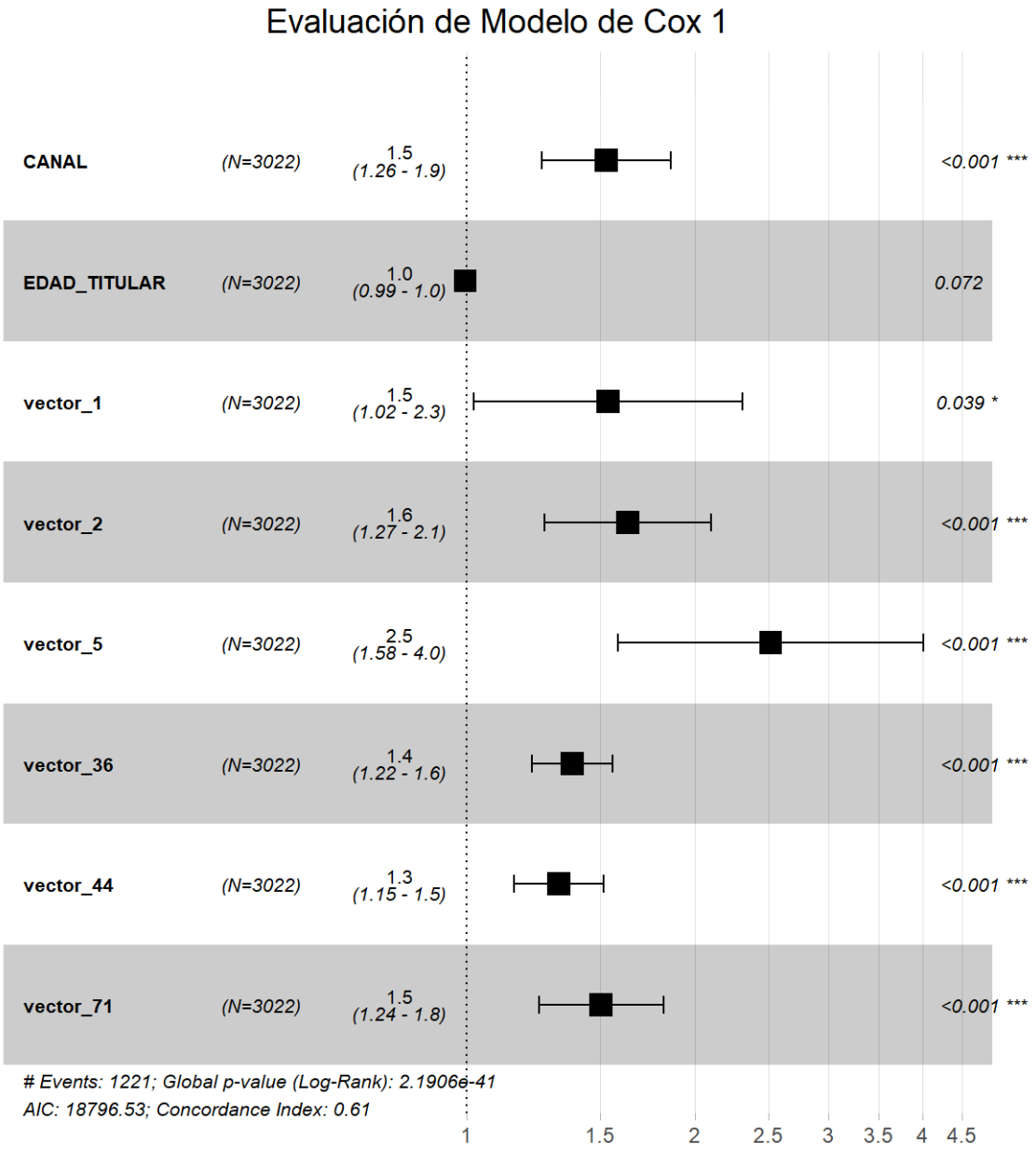


Figura 4.8: Evaluación de la exponencial de los coeficientes en la estimación del modelo de Cox para la población V1

principalmente a que los adultos mayores tienden a tener enfermedades crónicas como hipertensión y diabetes, esto haría que dar por terminado el seguro privado sea muy riesgoso para ellos, ya que sus gastos médicos pueden incrementarse con cualquier siniestro. Esto se complementa con la inclusión al modelo de la covariable EDAD_TITULAR que nos dice que a menor edad del titular mayor probabilidad de deserción lo que puede deberse a que los adultos de hoy son los llamados “millenials”, una generación bastante complicada que desea todo en ese momento, impacientes y de “gatillo rápido”; enfocarse en mantener a

este tipo de clientes satisfecho puede también ser la clave para reducir la deserción en el primer año.

Otra variable importante es el vector_2, lo que nos dice que el número de recortes que se le hace al cliente durante el primer año, influye en su decisión de desafiliarse. Si los recortes son mayores a 2 dólares los clientes tienen menor probabilidad de deserción. Los recortes se suelen dar principalmente porque el prestador que dio el servicio no tiene convenio con Humana S.A. y por ende sus precios son elevados, sin embargo, los afiliados los prefieren porque son prestadores de mucha confianza. Esto no lleva a la conclusión de que los clientes que tienen acceso a sus prestadores de confianza tienen menor probabilidad de deserción, inclusive si esto conlleva un mayor costo.

El vector_1 es otra de las variables influyente. Nos dice que si el valor presentado para pago al afiliado es mayor a 491 dólares en los últimos 6 meses entonces el riesgo de deserción es menor. Podríamos decir que a mayor uso del servicio menor riesgo de deserción pero solo parece ser cierto cuando el afiliado cobra sus reembolsos directamente, porque cuando es el prestador el que lo hace algo no va del todo bien, porque lo mismo no ocurre cuando tomamos en cuenta el valor presentado por parte del prestador. Investigar, cual puede ser la razón para que los afiliados que prefieren que se les pague el reembolso a ellos y no al prestador tienen menor riesgo de deserción puede ser una valiosa herramienta para

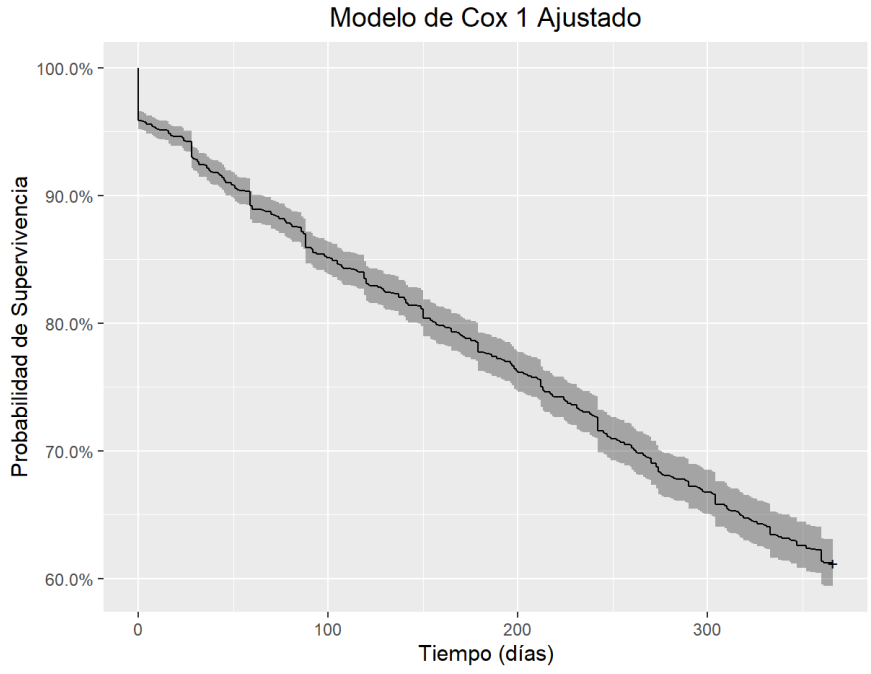


Figura 4.9: Estimador de función de Supervivencia del Modelo de Cox Ajustado para la población V1

entender el comportamiento respecto a la deserción.

La última covariable que analizaremos es el CANAL. El canal de venta directo (a través de asesores comerciales de Humana S.A.) parece tener mayor riesgo de deserción que otros canales como la venta mediante Brokers lo que es bastante sorprendente y un hallazgo poco intuitivo.

En base al modelo final ajustado, podemos construir la función de supervivencia de la población V1, que se puede ver en la figura 4.9.

4.5.9.2 Validación de hipótesis de riesgos proporcionales del modelo de Cox para la población V1

Para validar la hipótesis de riesgos proporcionales se utiliza la función “cox.zph”, incluida en el paquete “survival” [18], que evalúa los residuos de Schoenfeld y los p-valores de la hipótesis nula. En la tabla 4.11 se pueden ver los resultados.

Tabla 4.11: Prueba de Hipótesis de riesgos proporcionales de modelo de Cox para la población V1

	chisq	df	p
CANAL	0.323	1	0.570
EDAD_TITULAR	1.441	1	0.230
vector_1	0.459	1	0.498
vector_2	0.182	1	0.670
vector_5	0.329	1	0.566
vector_36	1.642	1	0.200
vector_44	2.979	1	0.084
vector_71	0.001	1	0.981
GLOBAL	8.683	8	0.370

Como podemos ver lo p-valores de todas las covariables superan 0,05 y también el p-valor global, de manera que aceptamos la hipótesis nula, por lo que podemos decir que el modelo cumple con la supuesto de riesgos proporcionales.

4.5.9.3 Estimación del modelo de Cox para la población V2

La estimación del modelo de Cox para la población V2, de la misma manera que con la población anterior, se inició con las variables que en el proceso de depuración se consideró que tenían buena capacidad de discriminación junto con todos los vectores, construidos

para esta población, que no se eliminaron. Las covariables que se tomaron en cuenta fueron las siguientes:

- ❒ EDAD_TITULAR
- ❒ NUMERO_AFILIADOS_MENORES
- ❒ TIEMPO_AFILIACION
- ❒ PRODUCTO_MH
- ❒ MORA_PROM_12M
- ❒ MORA_MAX_12M
- ❒ INCREMENTO_RELATIVO_2018
- ❒ INCREMENTO_ABSOLUTO_2018
- ❒ DIAS_FIN_CONTRATO
- ❒ SINIESTRALIDAD_12M
- ❒ PRIMA_PE
- ❒ INCREMENTO_ABS
- ❒ INCREMENTO_REL
- ❒ MORA_PROM_6M
- ❒ MORA_MAX_6M
- ❒ vector_2
- ❒ vector_27
- ❒ vector_30
- ❒ vector_37
- ❒ vector_50
- ❒ vector_52
- ❒ vector_55
- ❒ vector_61
- ❒ vector_68
- ❒ vector_72
- ❒ vector_73
- ❒ vector_75

Como puede verse se incluyó las variables MORA_PROM_6M y MORA_MAX_6M, debido a que en algunos análisis tuvieron excelentes resultados.

En base a las variables nombradas y a un arduo proceso de modelización se tuvo como resultado un modelo de riesgos proporcionales que cumple con todos los requisitos y supuestos. A continuación, se presenta el código de R y luego el resultado final:

Código de R 4.9: Código Modelo de Cox población V2

```
1 ###############################################################################
-
- ############## Modelo de Cox 2
-
5 Modelo_Cox2 <- coxph(formula = Surv(TIEMPO, CENSURA) ~ MORA_MAX_6M + vector_2  +
    vector_27 + vector_30 + vector_37 + vector_50 + vector_52 + vector_55 +
```

```
    vector_61 + vector_68 + vector_72 + vector_73 + vector_75,
    data = Base_Cox, y = TRUE, x = TRUE)

cox.zph(Modelo_Cox2)
```

En la tabla 4.12 podemos encontrar un resumen del modelo V2, entre los datos que podemos ver están los coeficientes de cada una de las covariables que ingresaron al modelo final junto con sus errores estándar y su significancia, los estadísticos para el contraste de hipótesis, el índice de concordancia (Concordance) y el R^2.

Los estadísticos de las prueba de contraste de hipótesis de Wald, razón de verosimilitud (LR) y Score tienen un p-valor por debajo de 0,01. En resumen, todas las pruebas de contraste nos dicen que el modelo está bien ajustado.

El modelo final tiene un índice de concordancia de 0,66 lo que es bueno, sin embargo, también tiene un R^2 poco prometedor (0,091). Basado en esto podemos decir que el modelo tiene poca capacidad predictiva, pero si nos puede dar una idea de las variables que influyen mayoritariamente en la deserción de clientes.

Antes de hacer un análisis del modelo, se explicará que es lo que mide cada covariable que ingresó al modelo:

1. **MORA_MAX_6M**. Máxima cantidad de días que el contrato estuvo en mora en los últimos 6 meses.
2. **vector_2**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando la suma total de los recortes es menor o igual a $61,61 y 1 (mayor riesgo de deserción) en caso contrario.
3. **vector_27**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando el número de días de mora máximo que tuvo el contrato en los últimos 6 meses fue de 6 y el incremento relativo en su última renovación no fue mayor a 15,6 % . Toma el valor de 1 (mayor riesgo de deserción) en caso contrario.
4. **vector_30**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando el titular del contrato es menor de 50 años y el contrato tiene 1484 días de afiliación como máximo. Toma el valor de 0 (menor riesgo de deserción) en caso contrario.
5. **vector_37**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando los recortes en los últimos 6 meses no superan los $120,00 y además la mora máxima

Tabla 4.12: Resumen del Modelos de Cox para la población V2

	Dependent variable:
	TIEMPO
MORA_MAX_6M	0.010***
	(0.003)
vector_2	0.349**
	(0.171)
vector_27	0.234***
	(0.078)
vector_30	0.382***
	(0.070)
vector_37	0.408***
	(0.108)
vector_50	0.239***
	(0.091)
vector_52	0.212***
	(0.063)
vector_55	0.407***
	(0.081)
vector_61	0.379***
	(0.098)
vector_68	0.195**
	(0.089)
vector_72	0.313**
	(0.127)
vector_73	0.478***
	(0.157)
vector_75	0.450***
	(0.122)
Observations	4,468
Concordance	0.663 (0.008)
R^2	0.091
Max. Possible R^2	0.986
Log Likelihood	−9,281.929
Wald Test	478.000*** (df = 13)
LR Test	425.758*** (df = 13)
Score (Logrank) Test	525.526*** (df = 13)
Note:	*p<0.1; **p<0.05; ***p<0.01

del contrato en los últimos 6 meses es menor o igual a 31 días. Toma el valor de 1 (mayor riesgo de deserción) en caso contrario.

6. **vector_50**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando la edad del titular es menor a 40 años y el incremento de primas de su última renovación fue mayor a 12,97 %. Toma el valor de 0 (menor riesgo de deserción) en caso contrario.

7. **vector_52**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando el producto es Metrohumana[15] y una mora máxima de 14 días en los últimos 3 meses. Toma el valor de 1 (mayor riesgo de deserción) en caso contrario.

8. **vector_55**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando valor presentado en los últimos 12 meses por diagnósticos crónicos y potencialmente crónicos es menor o igual a $328,62 y además el contrato tiene una siniestralidad en los últimos 12 meses de entre 11,6 % y 51,3 %. Toma el valor de 0 (menor riesgo de deserción) en caso contrario.

9. **vector_61**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando la siniestralidad del contrato en los últimos 12 meses es mayor a 51,3 % y el incremento de primas de su última renovación fue mayor a $6,72. Toma el valor de 0 (menor riesgo de deserción) en caso contrario.

10. **vector_68**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando el contrato no incluye afiliados de 2 años de edad o menos y su siniestralidad es menor o igual a 76,3 %. Toma el valor de 1 (mayor riesgo de deserción) en caso contrario.

11. **vector_72**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando el contrato incluye a al menos un afiliado menor de edad y su incremento de primas en su última renovación fue mayor a 15,6 %. Toma el valor de 0 (menor riesgo de deserción) en caso contrario.

12. **vector_73**. Variable binaria. Toma el valor de 0 (menor riesgo de deserción) cuando el valor presentado por diagnósticos crónicos es mayor a $619,28 y su siniestralidad en sus últimos 12 meses es menor o igual a 51,3 %. Toma el valor de 1 (mayor riesgo de deserción) en caso contrario.

[15]Son productos enfocados a los niveles socio-económicos 'medio alto'y 'alto'.

13. **vector_75**. Variable binaria. Toma el valor de 1 (mayor riesgo de deserción) cuando la edad del titular es de menos de 50 años y faltan más de 318 días para el final de la versión actual (es decir, que su última renovación fue reciente). Toma el valor de 0 (menor riesgo de deserción) en caso contrario.

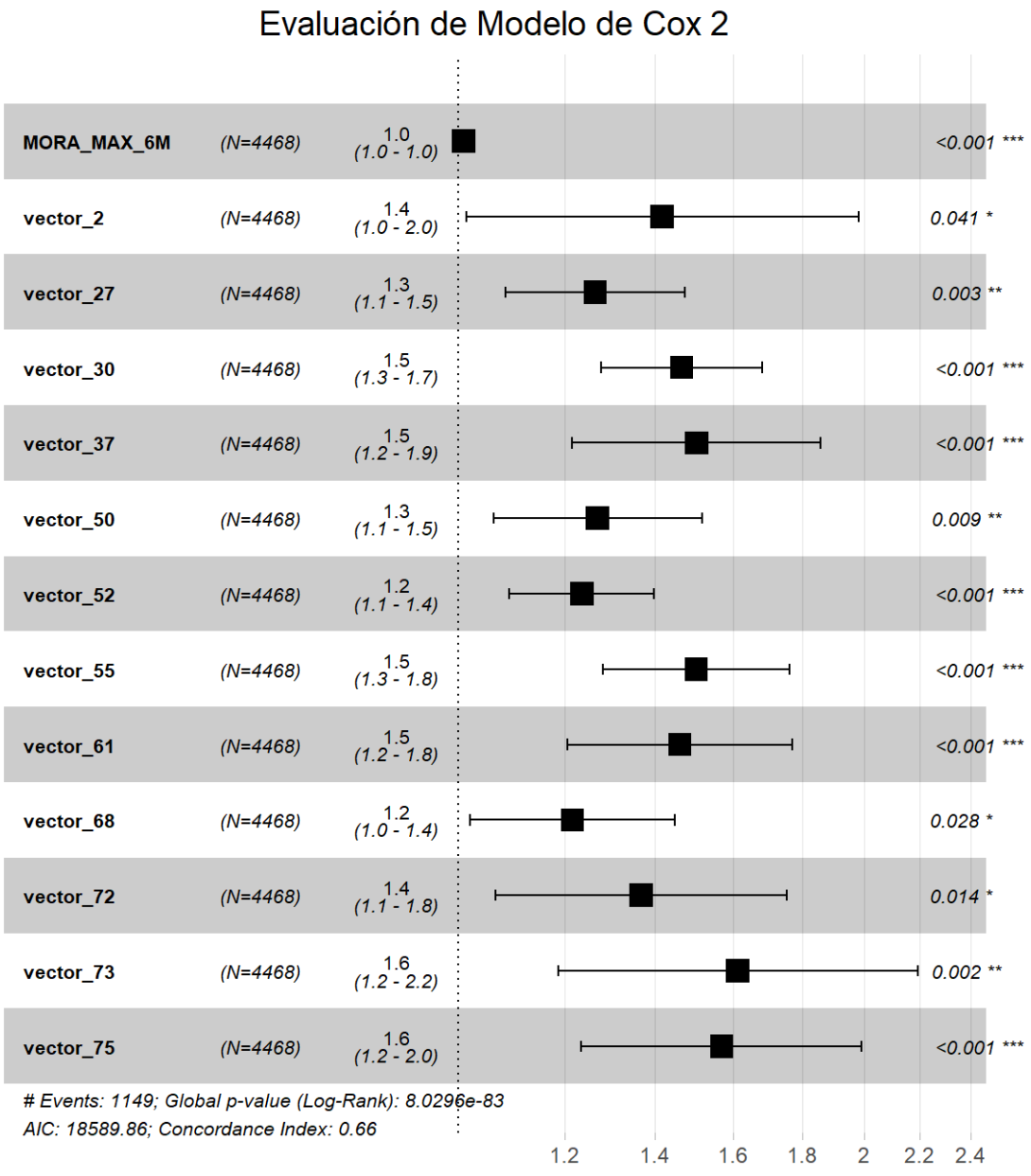


Figura 4.10: Evaluación de la exponencial de los coeficientes en la estimación del modelo de Cox para la población V2

También podemos ver la figura 4.10 que nos muestra el intervalo de confianza de la exponencial de los coeficientes del modelo ($\exp(\hat{\beta})$), poniéndolos en perspectiva unos con otros. Nos será útil para analizar cada covariable que ingresó al modelo.

La covariable que más peso tiene en el modelo es el vector_73, sin embargo, también tiene un error estándar muy alto. Esto nos dice que los clientes que tienen valores presentados por enfermedades crónicas pero que no superan una siniestralidad del 51 % tienen menor probabilidad de desertar. Aquí surge una pregunta clave ¿Por qué los clientes con siniestralidad baja tienen menor riesgo de desertar? ¿No debería ser lo contrario, que mientras más se usa el producto mayor es la lealtad hacia la empresa? Este comportamiento ha sido evidenciado tanto en el modelo como en los árboles de inferencia condicional escudriñados, por lo que podemos inferir que debe tratarse de algún problema que viene del servicio, un problema específico, que es necesario solucionar rápidamente. Regresando al análisis de la variable, vemos que las personas con enfermedades crónica pero que no usan el plan demasiado[16] tiene menor riesgo de desertar.

La segunda variable más importante del proceso viene a ser MORA_MAX_6M. Esto nos dice que mientras mayor sea su mora máxima en los últimos 6 meses mayor será su riesgo de deserción, lo que es bastante intuitivo principalmente porque es obvio que un contrato impago durante mucho tiempo debe tener poca importancia para el cliente, obviamente no es su prioridad y podría prescindir de él, y también porque la ley de medicina prepagada permite dar de baja a clientes por falta de pago durante un periodo de 3 meses.

También tenemos que el vector_75 es otra covariable influyente. Este predictor nos dice que si $EDAD\_TITULAR \leq 49$ y que $DIAS\_FIN\_CONTRATO > 318$ entonces el cliente tiene mayor probabilidad de deserción, lo que implica que estos clientes, mayoritariamente 'millennials'que acaban de pasar por una renovación de contratos y que probablemente tuvieron alguna inconformidad en ese proceso, tienen mayor riesgo de deserción. Como ya se dijo las personas que pertenecen a esta generación tienen actitudes muy particulares, investigar este fenómeno puede ayudar a entender a estos clientes y como retenerlos.

El vector_37 es otra de las variables importante. Nos dice que si el valor recortado en los últimos 6 meses es inferior a 120 dólares y la mora máxima en los últimos 12 meses es menor a 31 días entonces el cliente tiene menor probabilidad de desertar. Aquí surge un cambio de perspectiva, porque diferencia de la población V1, los clientes con mas de un año de afiliación prefieren que no se les haga recortes y si mantienen un comportamiento de pago medianamente adecuado esto nos dice que su riesgo de desertar es menor.

[16] la siniestralidad esperada en los planes individuales es de entre el 50 % al 60 %.

Comportamientos completamente opuestos entre la población V1 y V2 pueden deberse a la virtual falta de conocimiento por parte de los clientes en la primera población.

El vector_55 es un poco complicado porque nos dice que mientras un cliente hago un uso mínimo del plan (pero si haga uso de él) y sus enfermedades crónicas o potencialmente crónicas le generan gastos bajos o ningún gasto entonces el cliente tiene mayor riesgo de desertar. Nuevamente el problema parece apuntar hacia el servicio o los prestadores, algo en definitiva no está funcionando del todo bien al momento que los afiliados usan el servicio o asisten donde algún prestador.

El vector_61 nos devuelve al tema del servicio. A través de este vector vemos que los clientes con una siniestralidad mayor a 51 %, es decir que han usado mucho el servicio y que tuvo un incremento de al menos $6,72 tienen un mayor riesgo de deserción. La lectura de esto es que el afiliado piensa más o menos así “si estoy usando tu servicio y me encontré con muchos problemas y además de eso me incrementaste el precio del plan que contraté, sea justificado o no, sea mucho o poco, entonces prefiero no contar con este servicio”, algo que es bastante lógico.

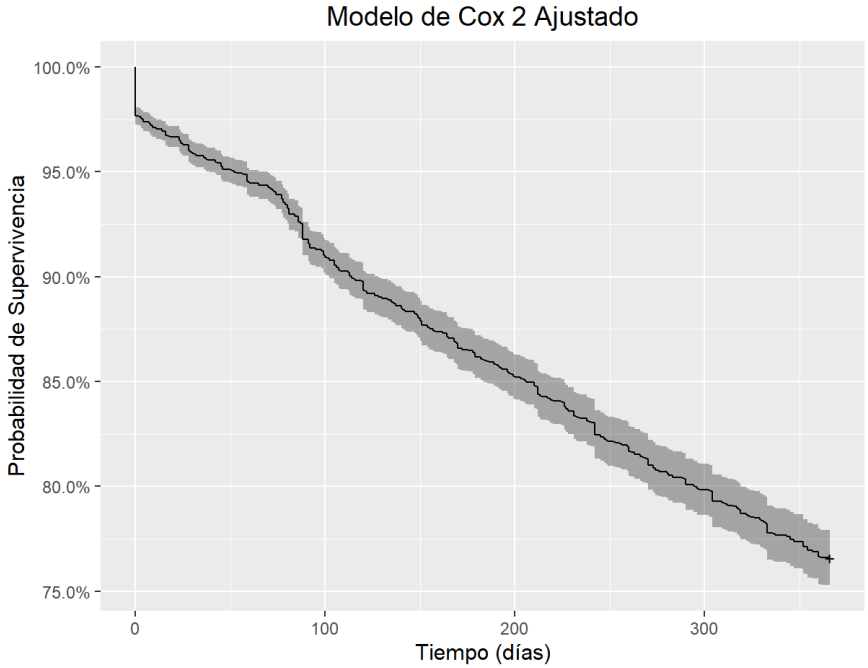


Figura 4.11: Estimador de función de Supervivencia del Modelo de Cox Ajustado para la población V2

El vector_30 nos dice que los contratos con titulares menores de 50 años y con 5 años o menos de afiliación tienen mayor riesgo de deserción. En pocas palabras, incluso cinco

años de afiliación no son suficientes para decir que un afiliado es fiel a la marca.

En base al modelo final ajustado, podemos construir la función de supervivencia de la población V2, que se puede ver en la figura 4.11.

4.5.9.4 Validación de hipótesis de riesgos proporcionales del modelo de Cox para la población V2

Para validar la hipótesis de riesgos proporcionales se utiliza la función “cox.zph”, incluida en el paquete “survival” [18], que evalúa los residuos de Schoenfeld y los p-valores de la hipótesis nula. En la tabla 4.13 se pueden ver los resultados.

Tabla 4.13: Prueba de Hipótesis de riesgos proporcionales de modelo de Cox para la población V2

	chisq	df	p
MORA_MAX_6M	1.312	1	0.252
vector_2	0.863	1	0.353
vector_27	0.334	1	0.563
vector_30	3.504	1	0.061
vector_37	0.267	1	0.605
vector_50	0.318	1	0.573
vector_52	1.333	1	0.248
vector_55	0.400	1	0.527
vector_61	2.023	1	0.155
vector_68	1.368	1	0.242
vector_72	0.044	1	0.835
vector_73	0.746	1	0.388
vector_75	1.786	1	0.181
GLOBAL	18.558	13	0.137

Como podemos ver lo p-valores de todas las covariables superan 0,05 y también el p-valor global, de manera que aceptamos la hipótesis nula, por lo que podemos decir que el modelo cumple con la supuesto de riesgos proporcionales.

4.5.10 VALIDACIÓN DE SUPUESTOS DE LOS MODELOS DE COX

4.5.10.1 Validación del supuesto de que las covariables continuas tienen una forma funcional adecuada

La verificación de este supuesto se lo realiza mediante los residuos de la martingala. No vamos a entrar en el cálculo matemático de estos residuos porque ya se tocó el tema anteriormente, pero sí vamos a comentar las propiedades de los residuos de la martingala. Evidentemente por definición el residuo de la martingala para un sujeto censurado será negativo. Para los sujetos no censurados el valor de los residuos puede ir desde -∞ hasta 1.

Lo ideal es que la linea que se traza en cada gráfico de estos residuos de la mantingala contra cada variable tienda al ajuste de una línea recta, lo que suele ser poco común pero en nuestro caso todo fue por buen camino. Ayala, Borges y Colmenares [52] recomiendan dicotomizar las covariables que no cumplen con este supuesto, pero que han demostrado tener la capacidad de discriminar a los sujetos en el estudio.

Además de lo anterior, se incluyó el gráfico de los residuos de la martingala contra cada vector creado en ambos modelos, lo que no es estrictamente necesario porque no se trata de variables continuas sino binarias. En estos gráficos en particular se colocaron dos líneas de color cían o verde (dependiendo de la población) que nos muestran la densidad de los datos en cada parte.

Las figuras 4.12 y 4.13 muestran las gráficas de los residuos de la martingala del modelo de riesgos proporcionales para la población V1 y en la primera figura esta la variable EDAD_TITULAR que es la única variable continua del modelo, la línea de tendencia en la gráfica de esta variable muestra claramente una recta.

Las figuras 4.14 a 4.17 muestran las gráficas de los residuos de la martingala del modelo de riesgos proporcionales para la población V2 y en la primera figura esta la variable MORA_MAX_6M que es la única variable continua del modelo, la línea de tendencia en la gráfica de esta variable muestra una desviación y no es tan clara una recta, sin embargo esta desviación se da cuando los datos son muy escasos y el intervalo de confianza se hace mucho más ancho, de manera que esta desviación no es suficiente para ir en contra

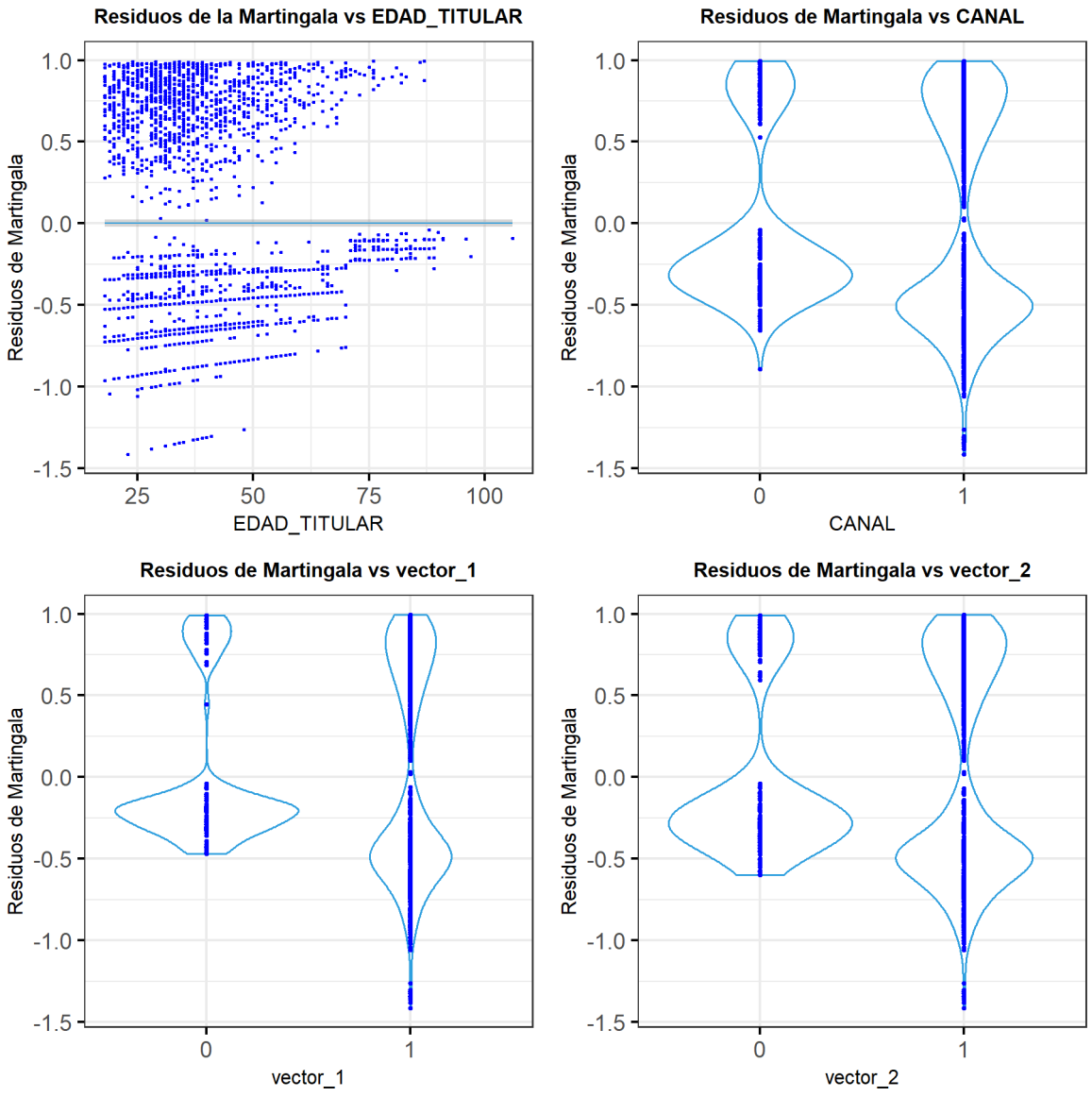


Figura 4.12: Gráficos de los residuos de la martingala contra las covariables que ingresaron al Modelo de Cox de la población V1 pag. 1 de 2

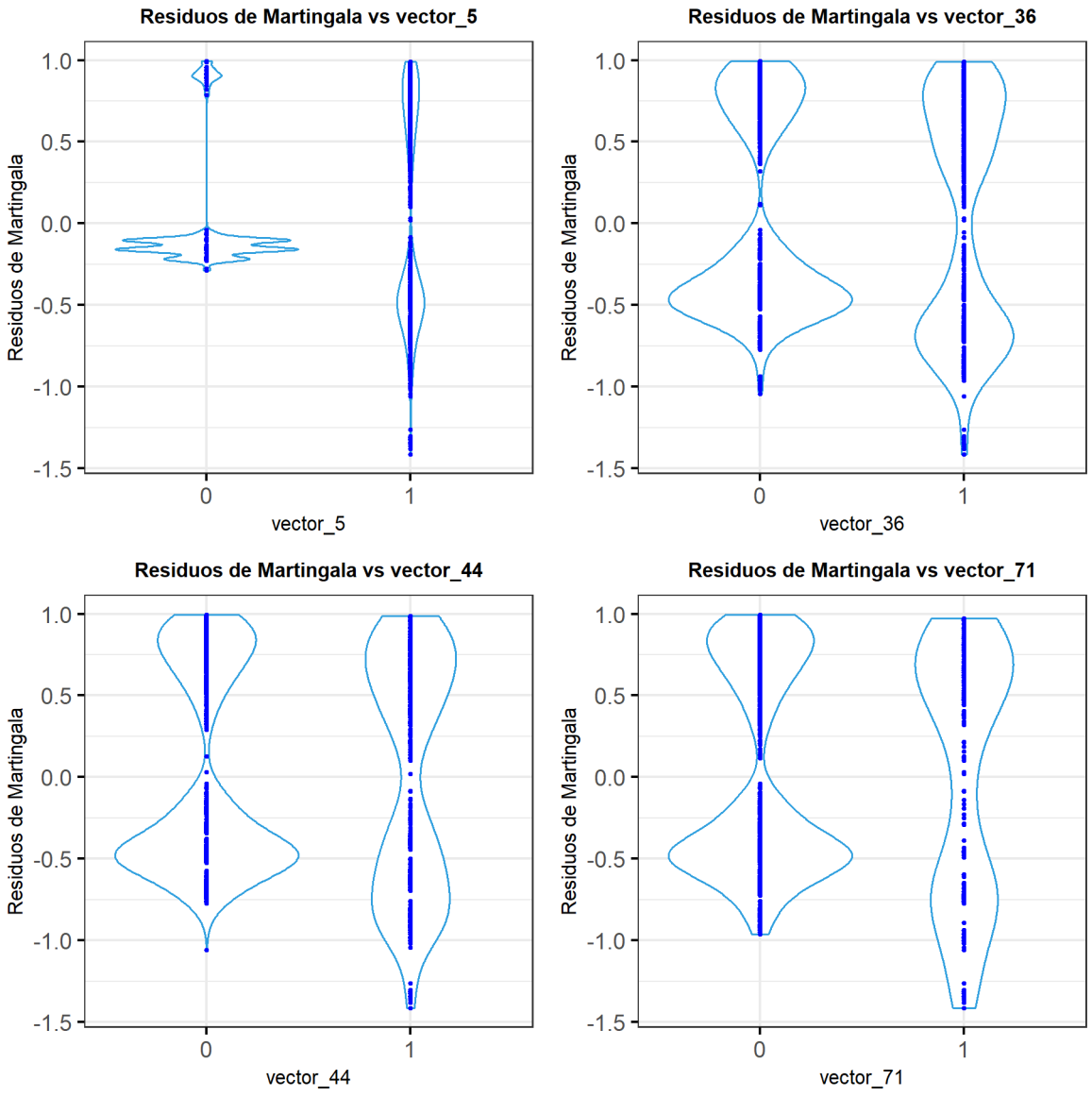


Figura 4.13: Gráficos de los residuos de la martingala contra las covariables que ingresaron al Modelo de Cox de la población V1 pag. 2 de 2

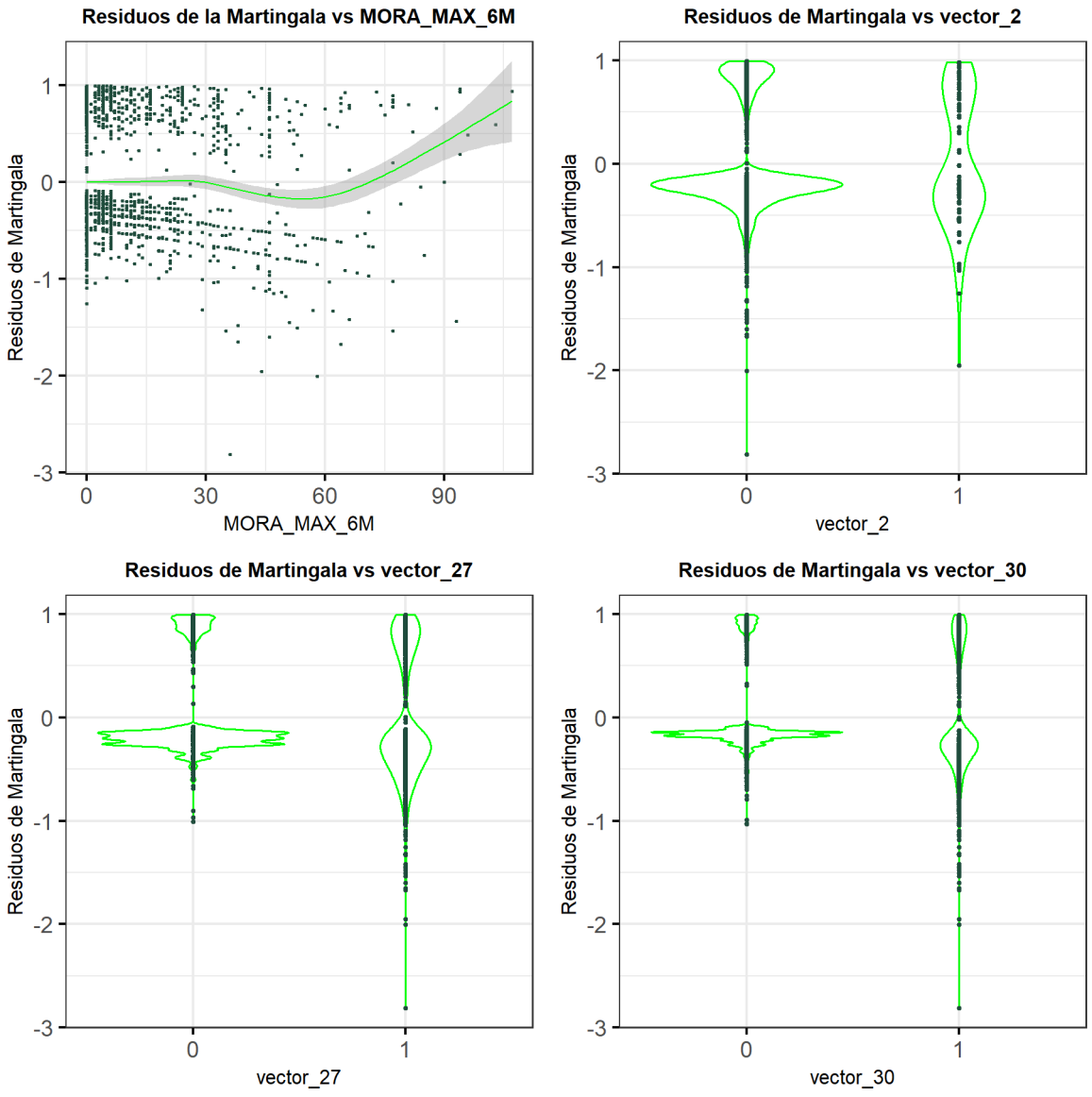


Figura 4.14: Gráficos de los residuos de la martingala contra las covariables que ingresaron al Modelo de Cox de la población V2 pag. 1 de 4

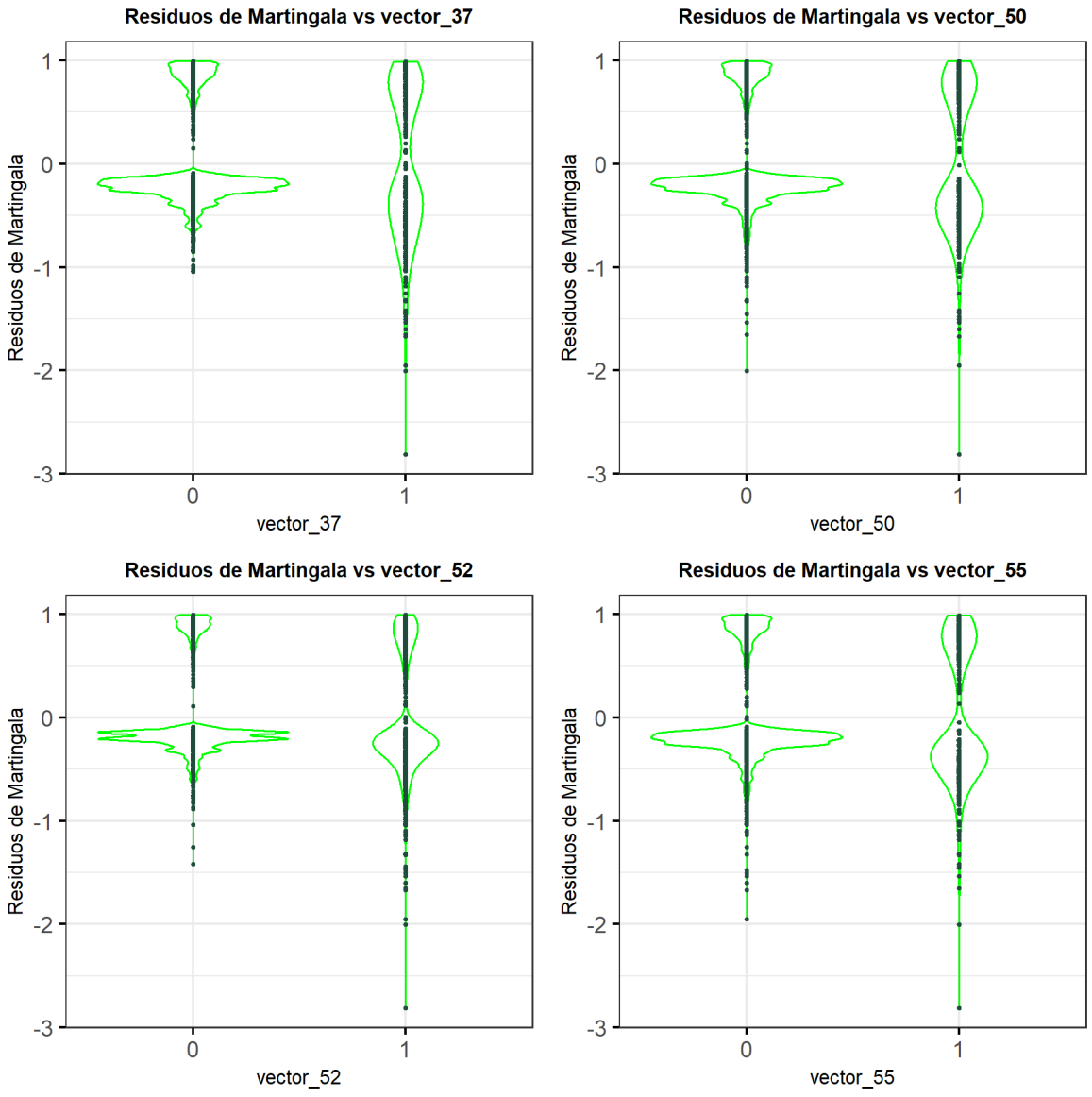


Figura 4.15: Gráficos de los residuos de la martingala contra las covariables que ingresaron al Modelo de Cox de la población V2 pag. 2 de 4

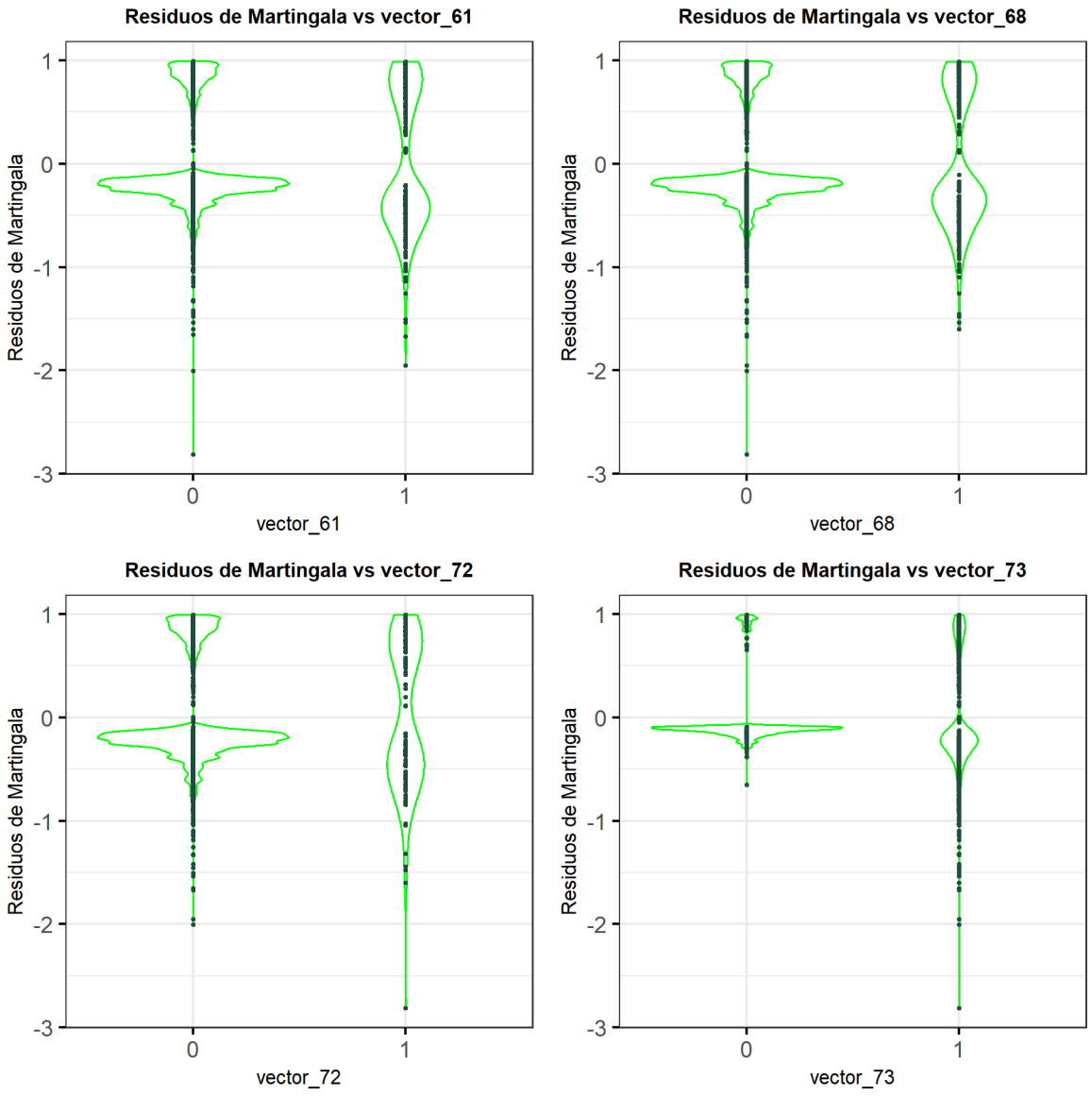


Figura 4.16: Gráficos de los residuos de la martingala contra las covariables que ingresaron al Modelo de Cox de la población V2 pag. 3 de 4

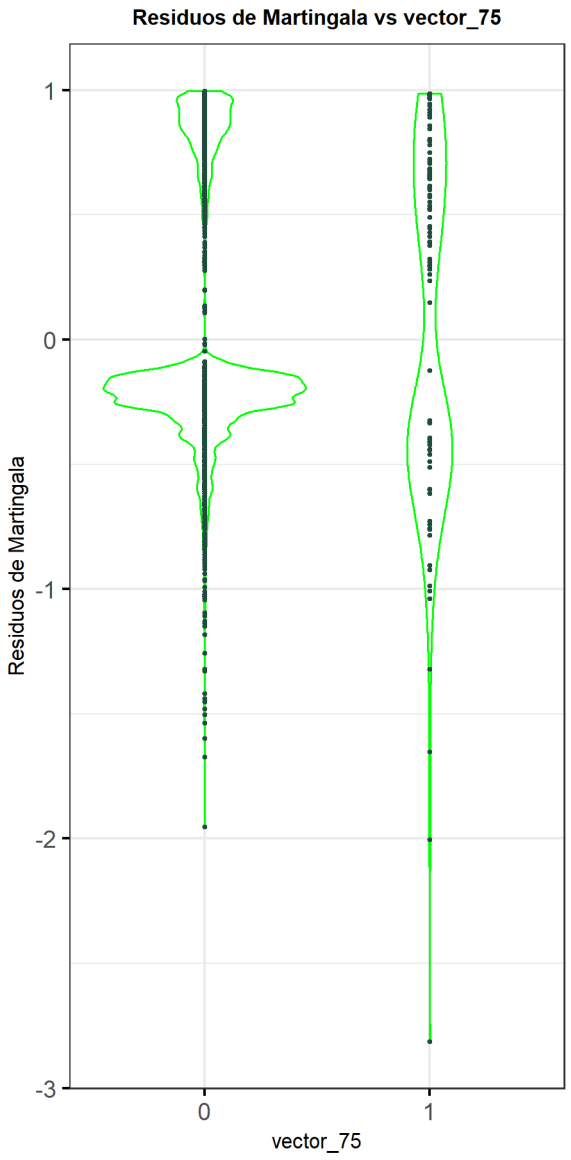


Figura 4.17: Gráficos de los residuos de la martingala contra las covariables que ingresaron al Modelo de Cox de la población V2 pag. 4 de 4

del supuesto que queremos corroborar.

Como vemos las covariables continuas de los dos modelos tienden a justar una línea recta, por lo que podemos decir que ambos modelos cumplen el supuesto.

4.5.10.2 Validación del supuesto de que los sujetos no tienen influencia en la estimación de cada coeficiente

La verificación de este supuesto se lo realiza mediante los residuos de la Puntajes (Score). Como se dijo en el capítulo 2 estos residuos tienen en cuanta la dependencia del tiempo de las covariables pero en nuestro caso todas las covariables de ambos modelos son independientes del tiempo.

Los residuos de puntajes se graficarán contra el id de cada observación (o sujeto). El id de la observación es el orden que se le dio al sujeto en la estimación del modelo (en función del riesgo y tiempo de fallo). De esta manera las primeras id corresponden a los sujetos censurados y luego los no censurados.

Lo ideal es que estos residuos de puntajes se acumulen alrededor de cero y que los residuos se ajusten a una recta en cero, es decir, que para ninguna covariable existan sujetos cuyos residuos de puntajes se alejen de cero y desvíen la recta en algún punto porque esto implicaría que este sujeto puede tener una influencia grande en la estimación del coeficiente de esa covariable. Es aconsejable poner particular énfasis en las covariables continuas.

La figura 4.18 muestra todos los gráficos de los residuos de puntajes para las covariables en el modelo de la población V1 y es claro que no existen sujetos que tengan influencia fuerte en la estimación de los coeficientes de ninguna de las covariables.

La figura 4.19 muestra todos los gráficos de los residuos de puntajes para las covariables en el modelo de la población V2 y como podemos ver no existen sujetos que tengan influencia fuerte en la estimación de los coeficientes de ninguna de las covariables.

Los ajustes de los datos de todas las covariables de los dos modelos forman una recta con pocos desvíos sobre cero y ningún sujeto tiene un residuo de puntaje alejado de cero en esos intervalos donde hay ligeros desvíos, por lo que podemos decir que ambos modelos cumplen el supuesto.

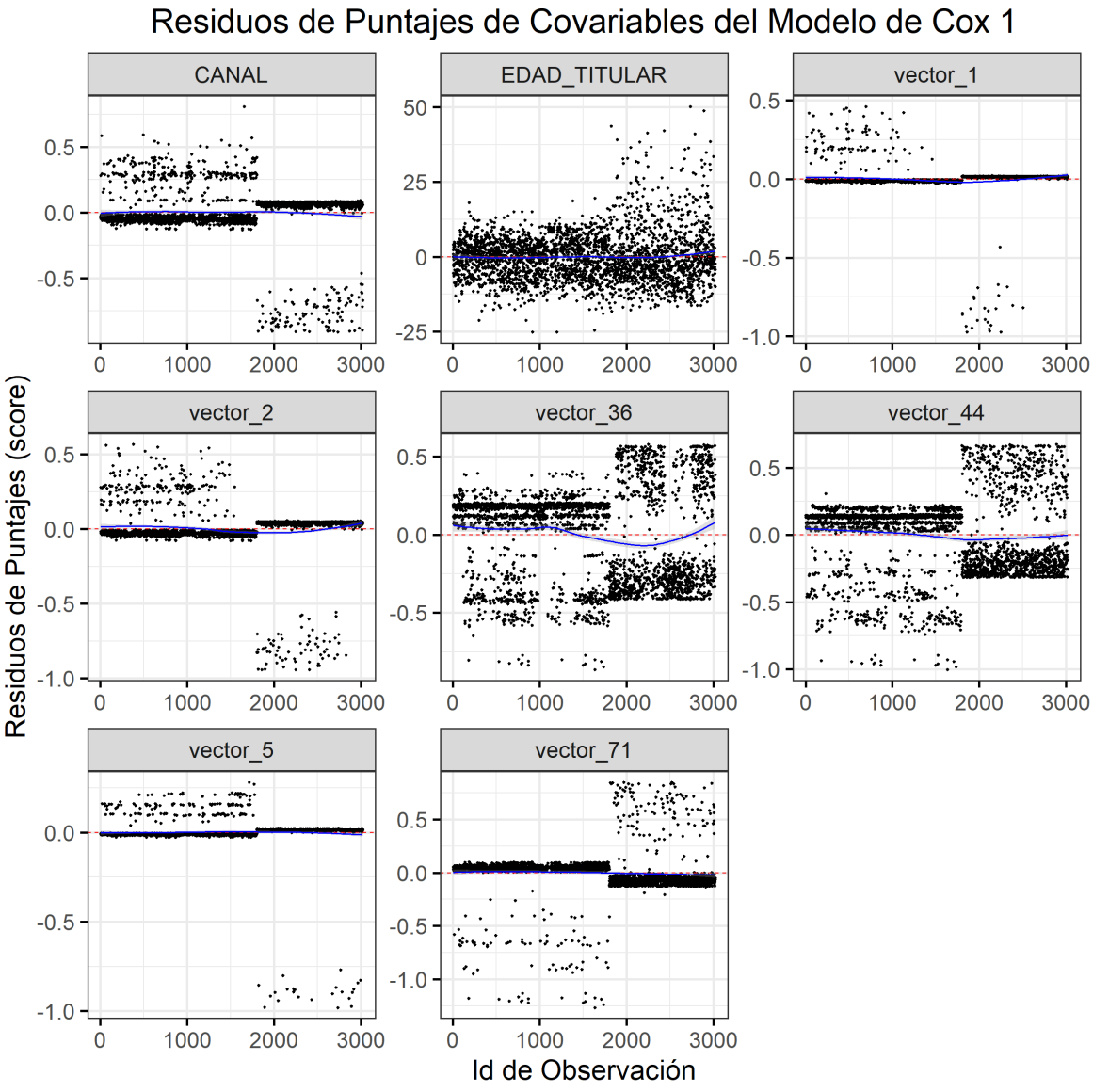


Figura 4.18: Residuos de Puntajes para cada covariables contra la id del sujeto en el modelo de Cox de la población V1

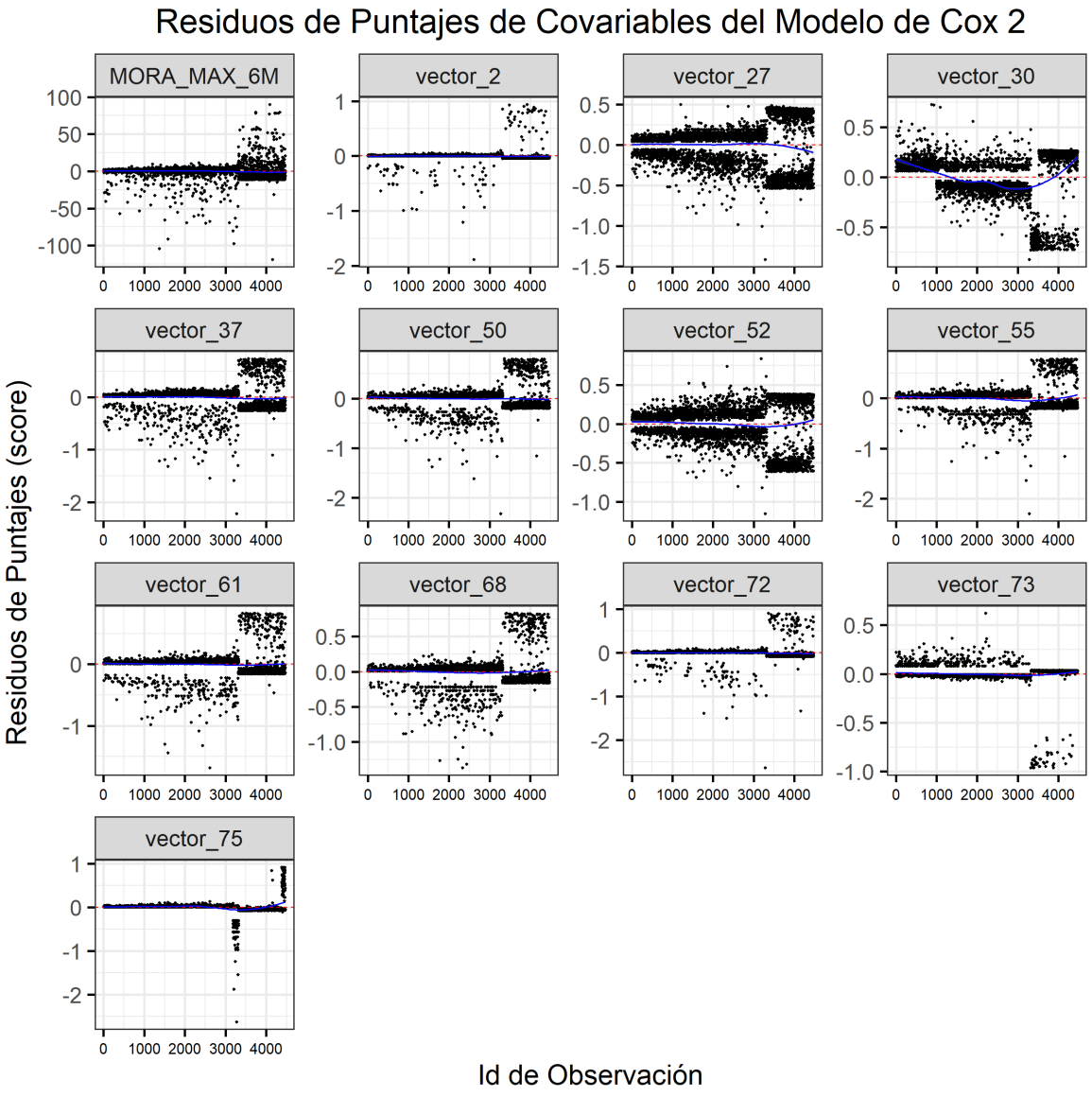


Figura 4.19: Residuos de Puntajes para cada covariables contra la id del sujeto en el modelo de Cox de la población V2

4.5.10.3 Validación del supuesto de que los sujetos no tienen influencia en la estimación del modelo

La verificación de este supuesto se lo realiza mediante los residuos de Desvíos (Desviance). Recordemos que la desviación de un modelo de Cox es el estadístico que se utiliza para cuantificar hasta qué punto el modelo actual que hemos estimado se aleja del modelo teórico que se ajustase perfectamente a nuestros datos [17]

Los residuos de desvíos se construyen transformando los residuos de la martingala de tal manera que produzcan valores simétricos en torno de 0, y para estos residuos de desvíos el rango de valores va desde $-\infty$ hasta $+\infty$. Aunque los residuos de desviación se distribuyen simétricamente en torno de cero si el modelo es adecuado, no tienen por qué sumar cero.

Los residuos de desvíos se graficarán contra la predicción lineal de cada sujeto, es decir, contra cada $\sum_{j=1}^{p} \hat{\beta}_j x_j$ donde i es el indicador del sujeto y j el de cada covariable que ingresó al modelo.

Lo ideal para estos residuos de desvíos es que ningún dato se aleje demasiado de cero y también es deseable que se ajusten a una recta que pasa por cero aunque esto no es del todo necesario.

La figura 4.20 nos muestra los residuos de desvíos graficados contra la predicción lineal de los sujetos en el modelo de Cox para la población V1. Puede verse que no hay datos que se alejen de cero.

La figura 4.21 nos muestra los residuos de desvíos graficados contra la predicción lineal de los sujetos en el modelo de Cox para la población V2. Como podemos ver no hay datos que se alejen de cero y tenga influencia en la estimación del modelo.

En conclusión, los valores de los residuos de desvíos de todos los sujetos en ambos modelos no se alejan demasiado de cero, por lo que podemos decir que ambos modelos cumplen el supuesto.

[17] denominado modelo completo o modelo saturado.

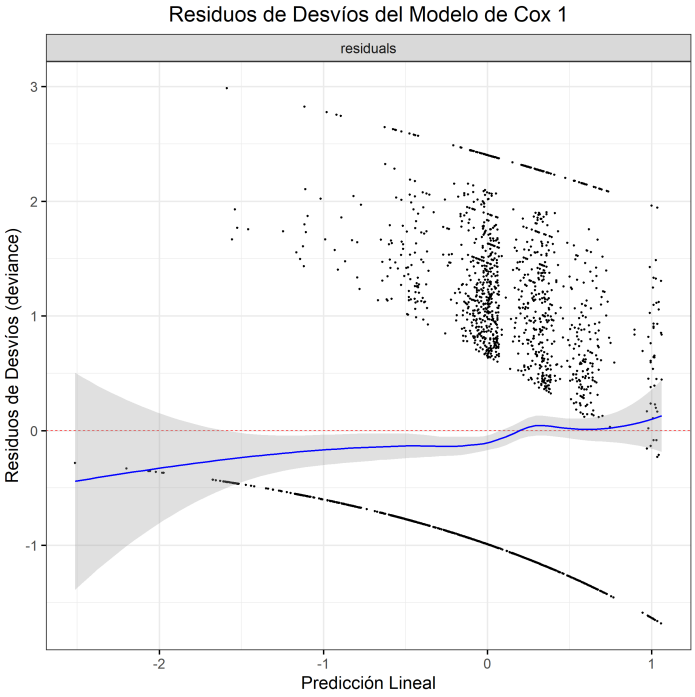


Figura 4.20: Residuos de Desvíos de las variables contra la predicción lineal en el modelo de Cox para la población V1

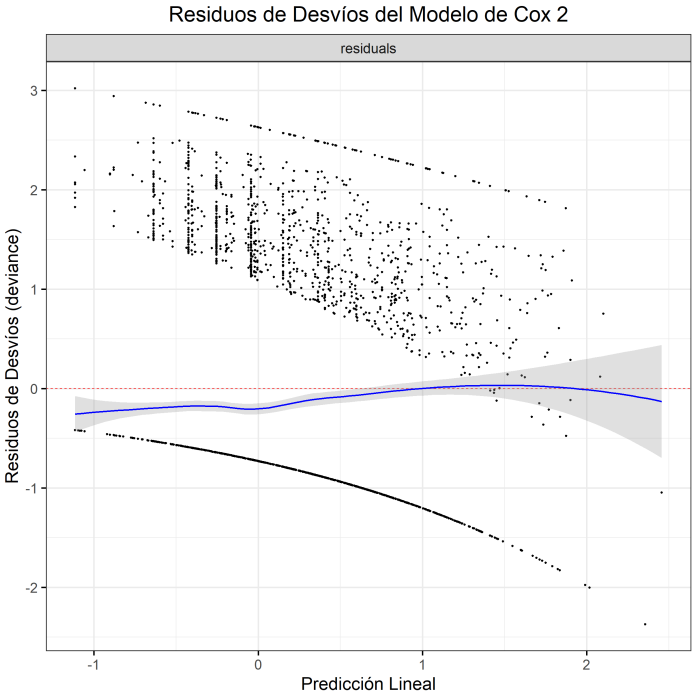


Figura 4.21: Residuos de Desvíos de las variables contra la predicción lineal en el modelo de Cox para la población V2

4.5.11 VALIDACIÓN EN BASES DE PRUEBA

Para la validación de los modelo con las bases de prueba usaremos el AUC tiempo dependiente y el Brier tiempo dependiente, que en prueba deberían dar resultados aceptables con AUC's de al menos 0,60 y Brier's que se alejen del Modelo Nulo (Null model) y en lo posible no superen 0,2.

Sería bueno comparar tanto el R^2 como el índice de concordancia en desarrollo y en prueba, pero debido a lo bajo del R^2 en desarrollo solo compararemos los índices de concordancia.

Para el modelo V1 el índice de concordancia en desarrollo fue de 0,609 (error estándar de 0,008) y el de prueba fue de 0,589, de manera que la diferencia entre ambos es de un poco más del doble del error estándar con lo que todavía podemos decir que el modelo no está sobreentrenado y es capaz de determinar las variables que pesan más en el evento de interés.

Para el modelo V2 el índice de concordancia en desarrollo fue de 0,663 (error estándar de 0,008) y el de prueba fue de 0,656, entonces como la diferencia es menor al error estándar podemos estar seguros de que el modelo no está sobreentrenado y es capaz de determinar las variables que pesan más en el evento de interés.

Para evaluar el AUC y Brier tiempo dependiente se uso la función "Score" del paquete riskReggression [26] del software R. Los cálculos se realizaron usando las bases de prueba de cada población respectivamente. Recordemos que estas bases son totalmente independientes de las bases de desarrollo con las que se creó cada modelos y fueron separadas desde un inicio con este fin.

El AUC y el Brier tiempo dependientes del modelo de cox 1 evaluado sobre la base de prueba de la población V1 pueden verse en las figuras 4.22 y 4.23 respectivamente. El AUC y el Brier tiempo dependientes del modelo de cox 2 evaluado sobre la base de prueba de la población V2 pueden verse en las figuras 4.24 y 4.24 respectivamente. En adelan cuando hablemos del AUC o Brier de alguno de los dos modelos se dará por sobreentendiendo que fueron evaluados sobre las bases de prueba de las poblaciones de los modelos en cuestión y que son tiempo dependientes.

El AUC del modelo de Cox de la población V1 muestra una tendencia creciente en el tiempo pero no llega a superar el 0,60 en ningún momento lo que no es bueno, esto nos dice

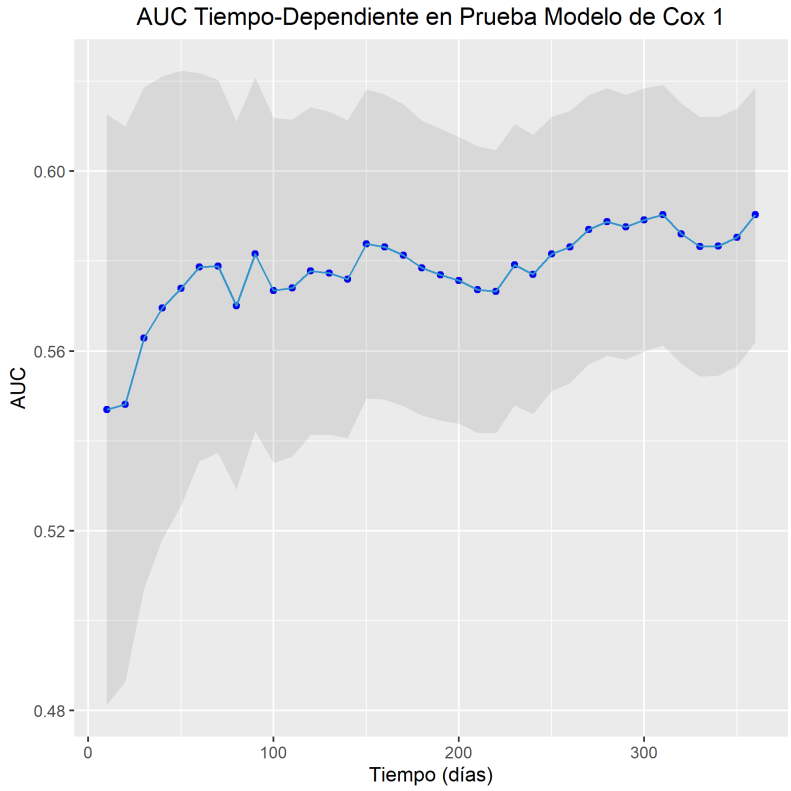


Figura 4.22: AUC tiempo-dependiente en Prueba para el modelo de riesgos proporcionales de la población V1

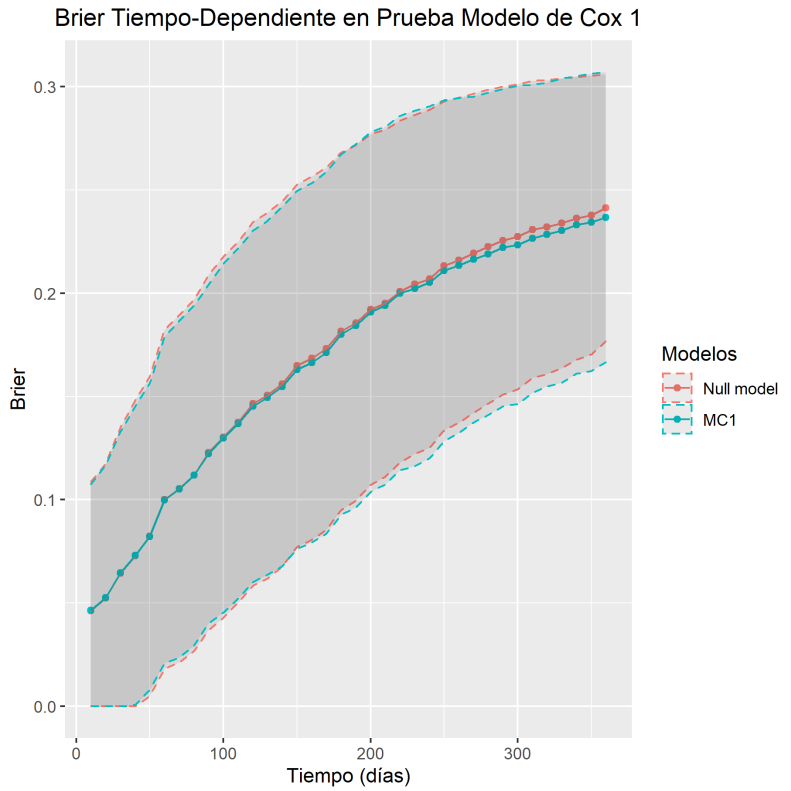


Figura 4.23: Brier tiempo-dependiente en Prueba para el modelo de riesgos proporcionales de la población V1

que el modelo necesita mejorarse. En cuanto al Brier del mismo modelo, este no parece despegarse del modelo nulo y además supera el 0,2 con relativa rapidez lo que nos dice que la capacidad predictiva del modelo no es muy buena (algo que ya sabíamos).

El AUC del modelo de Cox de la población V2 muestra mas bien una tendencia decreciente en el tiempo pero no llega a bajar de 0,65 en ningún momento lo que es bueno, esto nos dice que el modelo, lejos de su poca capacidad predictiva, si puede decirnos que variables influyen en la deserción de los clientes. En cuanto al Brier de este mismo modelo podemos decir que aunque no supera el 0,2 tampoco se separa lo suficiente del modelo nulo lo que muestra que la capacidad predictiva del modelo no es suficiente como para considerar su uso con ese fin.

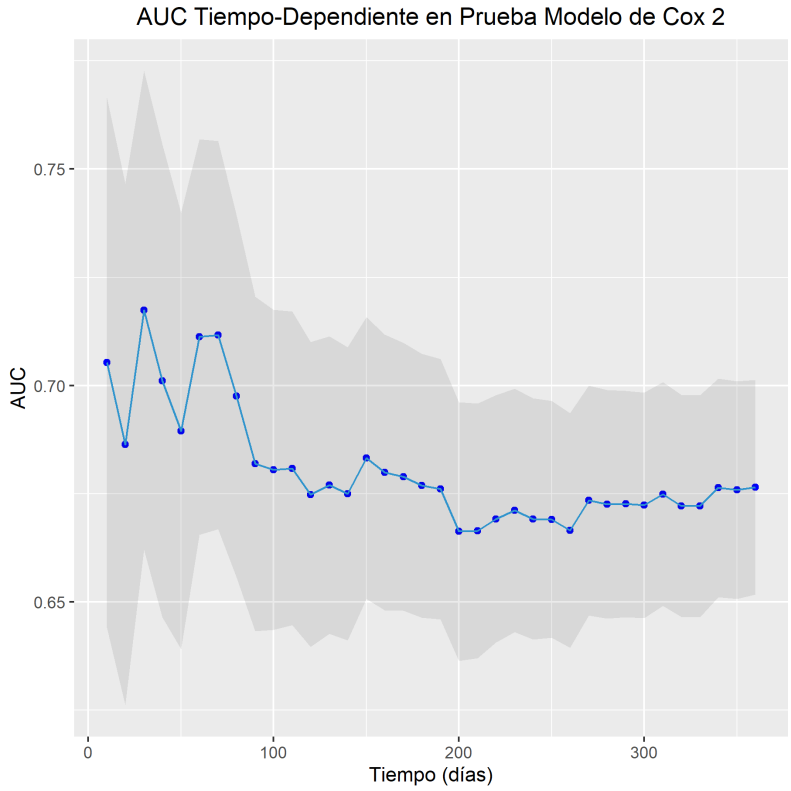


Figura 4.24: AUC tiempo-dependiente en Prueba para el modelo de riesgos proporcionales de la población V2

En base a lo anterior podemos decir que el modelo de Cox para la población V1 difícilmente nos será útil y debe ser mejorado, aunque si nos da luces de cuáles son los factores que influyen en la decisión de finalizar la relación contractual con la empresa en clientes con menos de un año de afiliación.

El modelo V2 nos lleva a excelentes conclusiones en cuanto a los factores que se relacionan con la deserción de clientes con más de un año de afiliación a la empresa pero todavía es

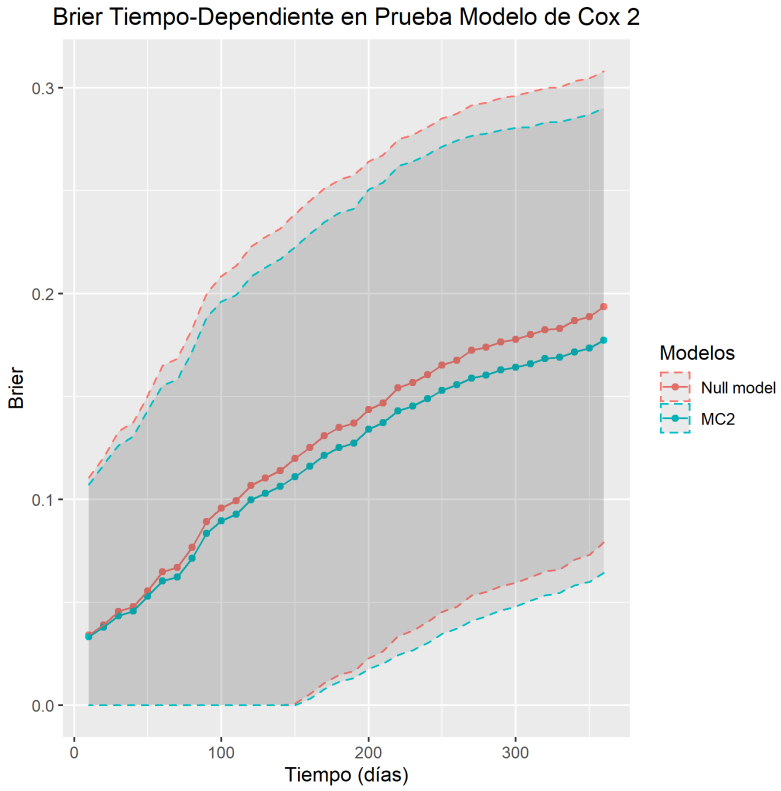


Figura 4.25: Brier tiempo-dependiente en Prueba para el modelo de riesgos proporcionales de la población V2

necesario complementar las variables que ya se tienen con otras que midan la calidad del servicio en distintos ámbitos.

La realidad es que parece ser que el primer año es una fase crítica en cuanto a deserción porque los clientes no saben que esperar y predecir su deserción se vuelve complicado, mientras que a partir del segundo año se puede tener una mejor predicción pero aún nos hace falta un eslabón o una luz que nos muestre el camino, ya que la variables relacionadas directamente con el servicio parecen ser la clave, pero no se pudo disponer de ellas.

5 CONCLUSIONES

- Aunque las variables que se obtuvieron fueron muchas (106), al parecer las variables estratégicas son aquellas un poco más esquivas y difíciles de medir como las que tienen que ver con el servicio. Las variables de servicio no están accesibles en la empresa Humana S.A. debido a que algunas de ellas no son capturadas de manera estricta. Es probable que teniendo estas variables se pueda construir modelos más asertivos.

- Debido a la constante aparición de el hecho de que a mayor siniestralidad parece incrementarse la probabilidad de deserción en los clientes de más de un año de afiliación, es posible que algún hecho aislado tocante al servicio o varios hechos interrelacionados también vinculados al servicio puede estar influyendo negativamente en el cliente. Este evento o eventos puede ocurrir desde la asistencia recibida por el prestador hasta el pago del reembolso. Sería importante para la empresa investigar al respecto.

- La edad del titular influye mucho en la deserción, al parecer los afiliados con edades de hasta 49 años, que tienden a ser menos pacientes, pueden estar tomando la decisión de terminar la relación contractual con la empresa debido a malentendidos o situaciones poco comunes. Se debe encontrar los principales puntos de dolor de estos clientes que aún sin usar el servicio pueden estar siendo afectados por eventos que se pueden prevenir.

- La empresa está perdiendo más clientes en el primer año de afiliación que en el resto del tiempo por lo que se presume que el inicio de la relación contractual es un punto crucial para la retención de los clientes. Es de vital importancia investigar al respecto debido a que muchas de las variables obtenidas no mostraron una clara relación con la deserción, por lo que es posible que la decisión de finalizar la relación contractual tenga más que ver con un tema subjetivo, aunque no puede descartarse que tenga

mucho que ver la mismas razones por las cuales los contratos con más de un año se desafilian.

- El AUC del modelo de riesgos proporcionales para la población V2 parecería empeorar mientras mayor es el tiempo, de lo que se deduce que es posible mejorar el modelo evitando tomar un periodo de tiempo demasiado largo, los datos mostraron que bastaría tomar uno, dos o hasta tres meses y se podría obtener mejores resultados inclusive si se usa otro modelo. En este caso balancear el número de datos censurados y no censurados sería clave, porque la cantidad de datos no censurados sería muy baja.

- Los afiliados con menos de un año requieren un seguimiento continuo para poder determinar que los hace desertar tan prontamente y que se puede hacer al respecto. Campañas agresivas de retención de clientes y medición de la satisfacción, junto con una adecuada capacitación al cliente pueden ayudar a reducir la caída de cartera. Todo esto debe acompañarse con una medición continua de resultados que luego pueda usarse para descubrir los puntos de dolor específicos que se deben atacar.

- La información actual de la empresa debe complementarse ya que es insuficiente para determinar las causas principales de la deserción, principalmente se debe empezar a registrar el momento exacto en el que el cliente toma la decisión de desafiliarse de la empresa, es decir, el momento de la entrega de la carta que, según la ley, el contratante debe hacer llegar a la empresa. Esto podría permitir desarrollar modelos a futuro que cuenten con información más cercana al instante de la decisión del cliente y por ende mucho más certera.

- Como ya se dijo el servicio es una pieza clave y buena parte de esa percepción de buen servicio depende de los prestadores, es decir, hospitales, clínicas, médicos, farmacias, centros de rehabilitación y más. Absorber las experiencias de cada prestador, evaluarlos y compartir estos datos sería otra herramienta poderosa. Por supuesto no estamos insinuando que se mida la deserción contra el uso del servicio de cada prestador (aunque no estaría de más).

- Los recortes también fueron otro punto negativo, al menos en los clientes con más de un año de afiliación, no tanto así en los clientes con un año o menos de relación contractual con la empresa. El hecho de que los resortes no afecten a los clientes que tienen un año o menos de afiliación nos muestra que es posibles que no estén del todo conscientes de qué son, cómo funcionan y porqué todas las empresas de medicina

prepagada se ven en la necesidad de usarlos. Evitar estos recortes sería valioso para la empresa, por supuesto no hablo de que se deje de hacerlos porque el riesgo de fraude está implícito, pero si se puede capacitar al cliente de una manera adecuada para que esté al tanto de este mecanismo universal, sepa las razones detrás del uso del mismo y que además comprenda que es preferible el uso de prestadores médicos acreditados y que tengan alianza con la empresa o que en todo caso sean verdaderamente confiables. Hay que tomar en cuenta que la ley orgánica que regula a las compañías que financien servicios de atención integral de salud prepagada hace corresponsable a la empresa de medicina prepagada de la atención médica recibida por el afiliado.

- El cambio de la prima de los contratos solo se puede hacerse de manera anual según la ley y usualmente las primas se incrementan debido a la inflación anual o a cambios de rango etario, es decir, que el cliente pasa a una edad donde sus gastos médicos esperados son mayores según los estudios actuariales aprobados por los entes de control. Estos incrementos no resultaron ser influyentes en la decisión de desafiliarse por parte del cliente o al menos no de manera directa. El incremento de precios parece ser más bien un complemento secundario a la hora de que el cliente decide y en la mayoría de los casos no es determinante.

6 REFERENCIAS BIBLIOGRÁFICAS

[1] D. Aaker y A. Biel, *Advertising and consumer psychology. Brand equity & advertising: Advertising's role in building strong brands*. Lawrence Erlbaum Associates, 1993.

[2] P. Kotler y K. L. Keller, *Dirección de Marketing*. Pearson Prentice Hall, 2006.

[3] F. F. Reichheld y T. Teal, *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value*. Harvard Business School Press, 1996.

[4] T. van Vuuren, M. Roberts-Lombard y E. van Tonder, «Customer satisfaction, trust and commitment as predictors of customer loyalty within an optometric practice environment», *Southern African Business Review*, págs. 81-96, 2012.

[5] S. Dillehay, «Ways to improve patient loyalty», *Review of Optometry*, pág. 8, 2006.

[6] J. P. Klein y M. L. Moeschberger, *SURVIVAL ANALYSIS Techniques for Censored and Truncated Data*. Springer, 1997.

[7] E. L. Kaplan y P. Meier, «Nonparametric estimation from incomplete observations», *Journal of the American Statistical Association*, vol. 53, n.º 282, págs. 458-481, jun. de 1958.

[8] N. Mantel, «Evaluation of survival data and two new rank order statistics arising in its consideration», *Cancer Chemotherapy Reports*, vol. 50, n.º 3, págs. 163-170, mar. de 1966.

[9] E. A. Gehan, «A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples», *Biometrika*, vol. 52, n.º 1/2, págs. 203-223, jun. de 1965.

[10] N. Breslow, «A Generalized Kruskal-Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship», *Biometrika*, vol. 57, n.º 3, págs. 579-594, dic. de 1970.

[11] R. E. Tarone y J. Ware, «On Distribution-Free Tests for Equality of Survival Distributions», *Biometrika*, vol. 64, n.º 1, págs. 156-160, abr. de 1977.

[12] R. Peto y J. Peto, «Asymptotically Efficient Rank Invariant Test Procedures», *Journal of the Royal Statistical Society A*, vol. 135, n.º 2, págs. 185-207, 1972.

[13] P. K. Andersen, O. Borgan, R. Gill y N. Keiding, «Linear nonparametric tests for comparison of counting processes, with applications to censored survival data (with discussion)», *International Statistical Review*, vol. 50, n.º 3, págs. 219-244, dic. de 1982.

[14] T. R. Fleming y D. P. Harrington, «A class of hypothesis tests for one and two sample censored survival data», *Communications In Statistics*, vol. 10, n.º 8, págs. 763-794, 1981.

[15] D. R. Cox, «Regression Models and Life Tables», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, págs. 187-220, 2 1972.

[16] P. Allison, «Event History Analysis», *Sage University Paper*, 46 1984.

[17] D. R. Cox, «Partial Likelihood», *Biometrika*, vol. 62, págs. 269-276, 2 1975.

[18] T. M. Therneau, *A Package for Survival Analysis in R*, R package version 3.2-3, 2020. dirección: https://CRAN.R-project.org/package=survival.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. dirección: http://www.R-project.org/.

[20] N. Breslow, «Covariance Analysis of Survival Data under the Proportional Hazards Model», *International Statistical Review*, vol. 43, págs. 43-54, 1974.

[21] B. Efron, «The Efficiency of Cox's Likelihood Function for Censored Data», *Journal of the American Statistical Association*, vol. 72, págs. 557-565, 1977.

[22] E. T. Lee y W. J. Wenyu, *Satistical Methods for suvirval data analysis*. Oklahoma City: Wiley Series, 2003, págs. 319-322.

[23] T. Therneau y P. Grambsch, *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, 2000.

[24] F. Harrel y K. Lee, «Verifying assumptions of the Cox proportional hazards model», en *Proceedings of the Eleventh Annual SAS Users Group International Conference*, Atlanta, Georgia: SAS Institute, 1986, págs. 823-828.

[25] P. Blanche, C. Proust-Lima, L. Loubere, C. Berr, J. F. Dartigues y H. Jacqmin-Gadda, «Quantifying and Comparing Dynamic Predictive Accuracy of Joint Models for Longitudinal Marker and Time-to-Event in Presence of Censoring and Competing Risks», *Biometrics*, vol. 71, n.º 1, págs. 102-113, 2015.

[26] T. A. Gerds y B. Ozenne, *riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks*, R package version 2020.02.05, 2020. dirección: https://CRAN.R-project.org/package=riskRegression.

[27] G. Brier, «Verification of Forecasts Expressed in Terms of Probability», *Monthly Weather Review*, vol. 78, n.º 1, págs. 1-3, abr. de 1950.

[28] Y. Zheng, T. Cai, Y. Jin y Z. Feng, «Evaluating prognostic accuracy of biomarkers under competing risk», *Biometrics*, vol. 68, n.º 2, págs. 388-396, jun. de 2012.

[29] P. Blanche, J.-F. Dartiguesa y H. Jacqmin-Gaddaa, «Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks», *Statistics in Medicine*, vol. 32, págs. 5381-5397, 2013.

[30] L. Parast, S.-C. Cheng y T. Cai, «Landmark Prediction of Long Term Survival Incorporating Short Term Event Time Information», *Journal of the American Statistical Association*, vol. 107, n.º 500, págs. 1492-1501, 2012.

[31] R. Schoop, J. Beyersmann, M. Schumacher y H. Binder, «Quantifying the predictive accuracy of time-to-event models in the presence of competing risks», *Biometrical Journal*, vol. 53, n.º 1, págs. 88-112, feb. de 2011.

[32] R. Schoop, E. Graf y M. Schumacher, «Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates», *Biometrics*, vol. 64, n.º 2, págs. 603-610, jun. de 2008.

[33] H. Hung y C. Chiang, «Estimation methods for time-dependent AUC models with survival data.», *Canadian Journal of Statistics*, vol. 38, n.º 1, págs. 8-26, mar. de 2010.

[34] J. Mingers, «Expert Systems—Rule Induction with Statistical Data», *The Journal of the Operational Research Society*, vol. 38, n.º 1, págs. 39-47, ene. de 1987.

[35] T. Hothorn, K. Hornik y A. Zeileis, «Unbiased Recursive Partitioning: A Conditional Inference Framework», *Journal of Computational and Graphical Statistics*, vol. 15, n.º 3, págs. 651-674, 2006.

[36] H. Strasser y C. Weber, «On the Asymptotic Theory of Permutation Statistics», *Mathematical Methods of Statistics*, vol. 8, págs. 220-250, 1999.

[37] P. Westfall y S. Young, *Resampling-Based Multiple Testing*. New York: John Wiley & Sons, 1993.

[38] L. Breiman, J. Friedman, R. Olshen y C. J. Stone, «Classification and Regression Trees», 1984.

[39] Y.-S. Shih, «Families of splitting criteria for classification trees», *Statistics and Computing*, vol. 9, págs. 309-315, 1999.

[40] S. M. O'Brien, «Cutpoint Selection for Categorizing a Continuous Predictor», *Biometrics*, vol. 60, n.º 2, págs. 504-509, jun. de 2004.

[41] E. Redacción, *TOP 5 Medicina prepagada*, https://www.ekosnegocios.com/, jun. de 2018.

[42] M. Masarifoglu y A. H. Buyuklu, «Applying Survival Analysis to Telecom Churn Data», *American Journal of Theoretical and Applied Statistics*, vol. 8, n.º 6, págs. 261-275, nov. de 2019.

[43] V. Balboa, «Modelos de predicción clínica en estudios de supervivencia Comparación de dos clasificaciones en el pronóstico de pacientes con síndrome mielodisplásico», Tésis de Maestría, Universidade de Santiago de Compostela, 2013.

[44] D. Van den Poel y B. Lariviere, «Customer attrition analysis for financial services using proportional hazard models», *European Journal of Operational Research*, vol. 157, n.º 1, págs. 196-217, 2004.

[45] NewVoiceMedia, *Customers Are Fighting Back against Poor Customer Service*, http://www.newvoicemedia.com, mayo de 2013.

[46] S. Ross, *Introduction to Probability Models*, Décima. Elsevier, 2010.

[47] J. H. Camelo Soares, J. L. Nascimento Barbosa, L. Araujo Lopes, G. Veras Magalhaes Junior, R. de Andrade Lira Rabelo, E. Baptista Passos y P. de Alcantara dos Santos Neto, «How to Avoid Customer Churn in Health Insurance/Plans? A Machine Learn Approach», *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, págs. 559-562, 2019.

[48] J. Su, K. Cooper, T. Robinson y B. Jordan, *Customer Retention Predictive Modeling in HealthCare Insurance Industry*. BlueCross BlueShield, 2010.

[49] C. Huigevoort, «Customer churn prediction for an insurance company», Master Thesis, Eindhoven University of Technology, abr. de 2015.

[50] V. Bhandari y P. Boutros, «Comparing continuous and discrete analyses of breast cancer survival information», *Genomics*, vol. 108, n.º 2, págs. 78-83, ago. de 2016.

[51] T. Hothorn y A. Zeileis, «partykit: A Modular Toolkit for Recursive Partytioning in R», *Journal of Machine Learning Research*, vol. 16, págs. 3905-3909, 2015. dirección: http://jmlr.org/papers/v16/hothorn15a.html.

[52] M. A. Ayala, R. E. Borges y G. Colmenares, «Análisis de supervivencia aplicado a la banca comercial venezolana», *Revista Colombiana de Estadística*, vol. 30, n.º 1, págs. 97-113, jun. de 2007.

7 ANEXOS

7.1 COEFICIENTE DE CORRELACIÓN POR RANGOS DE SPEARMAN

El coeficiente de correlación por rangos de Spearman es un estadístico no paramétrico de gran utilidad en aquellos análisis de datos en donde se desea conocer la relación lineal entre variables cuyas escalas de medidas sean al menos ordinales, o que exista suficientes evidencias de que las variables en estudio a pesar de ser cuantitativas no siguen una distribución normal.

El coeficiente de correlación de Spearman se define como: Sean X e Y dos variables aleatorias cualitativas ordinales tomadas de una muestra de tamaño n, con categorías A_i y B_i , y sean x_i e y_i los rangos que les corresponden a A_i y B_i, entonces

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Donde $d_i = x_i - y_i$ para todo $i = 1, 2, ..., n$.

El estadístico r_s puede tomar valores entre -1 y 1, igual que el coeficiente de correlación de Pearson y su interpretación también es idéntica.

La prueba de significancia de r_s trata de probar la hipótesis nula H_0: $No\ existe\ asociación\ entre\ las\ variables$ contra una hipótesis alternativa bidireccional o unidireccional.

7.2 DISTANCIA DE MAHALANOBIS

En estadística, la Distancia de Mahalanobis es una medida de distancia introducida por Mahalanobis en 1936. Su utilidad radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales. Se diferencia de la distancia euclídea en que tiene en cuenta la correlación entre las variables aleatorias.

La distancia de Mahalanobis es una de las alternativas más usadas para identificar y depurar datos atípicos multivariantes. Para calcular la distancia de Mahalanobis tenemos

$$d_m(\vec{x_1}, \vec{x_2}) = \sqrt{(\vec{x_1} - \vec{x_2})^T \Sigma^{-1} (\vec{x_1} - \vec{x_2})}$$

Donde, Σ es la matriz de covarianzas.

La distancia de Mahalanobis sigue aproximadamanete una χ^2 con n grados de libertad, donde n es la dimensión de las variables aleatorias.

Se usará el software R, para calcular esta distancia. R tiene varias funciones en distintos paquetes que permiten el cálculo de la distancia de Mahalanobis, en este trabajo usamos la función “mahalanobis” del paquete “stats” que viene por defecto en R.

La función mahalanobis exige al menos tres argumentos **x**, **center** y **cov**. El argumento **x** es una matriz de $p \times q$ tal que p es el número de variables y q, en nuestro caso, es el número de contratos. Básicamente, **x** es la matriz que contiene la información de cada variable por contrato.

El argumento **center** contiene los centros que hayamos decidido usar, en nuestro caso usaremos la media de cada variable y cada contrato será medido contra estas.

El argumento **cov** es la matriz de covarianza $(p \times p)$ de las variables.

Siendo x la matriz antes descrita, usamos en R el siguiente script:

Código de R 7.1: Código R para distancia de Mahalanobis

```
> Sx <- cov(x)
> D2 <- mahalanobis(x, colMeans(x), Sx)
> p <- pchisq(D2,ncol(x),lower.tail=FALSE)
```

Este script nos proporciona tres cosas **D2** que es la variable que contiene el cálculo de la distancia de Mahalanobis y **p** que es el vector de los p-valores.

7.3 DICCIONARIO DE TÉRMINOS

Atención hospitalaria.- Atención médica que requiere hospitalización de la persona.

Atención ambulatoria.- Atención médica que no requiere hospitalización de la persona.

Atención emergencia.- Atención médica que debe darse de manera inmediata, normalmente a través del área de emergencias de un hospital, y que puede derivar en una hospitalización o no.

Carencia.- Periodo de tiempo que debe transcurrir antes de que el afiliado tenga contractualmente derecho a la cobertura de un tipo de atención médica en especifico.

Copago.- contractualmente la empresa se obliga a pagar una parte de los gastos médicos de los afiliados en caso de un siniestro y la otra parte, la que paga el afiliado, es llamado copago.

Deducible.- Monto que se deduce del valor pagado neto por una única vez y hasta cubrir el total del deducible.

Enfermedad Aguda.- Son aquella que tiene un comienzo súbito y una evolución rápida. El periodo de convalecencia suele ser corto.

Enfermedad Crónica.- son enfermedades que suelen durar mucho tiempo e inclusive durante toda la vida del paciente. Suelen generar gastos continuos durante un largo periodo de tiempo.

Enfermedad Potencialmente Crónica.- son enfermedades que pueden derivar en casos crónicos o no, dependiendo de las circunstancias particulares de cada paciente.

Enfermedades preexistentes.- Son aquellas enfermedades, patologías o discapacidades, usualmente crónicas, existentes en el afiliado o beneficiario antes del momento de la contratación de un plan de medicina prepagada.

Facturación.- Es el valor total de las primas pagadas por el afiliado.

Pago al afiliado.- Es una modalidad de pago de reembolsos, en la que el afiliado presenta facturas por gastos médicos a la empresa y ésta le devuelve los valores pagados según las obligaciones contractuales.

Periodo de carencia.- Periodo de tiempo que el afiliado debe esperar desde que contrato

el plan de medicina prepagada hasta que puede hacer uso de algún beneficio en particular.

Prima.- Es el valor que paga el afiliado por los servicios de la empresa, puede ser mensual, trimestral o hasta anual.

Recorte.- Cuando la empresa de medicina prepagada detecta que los valores presentados por algún tratamiento, consulta o medicamento están muy por encima de los valores de mercado, se recortan estos valores, es decir se los reduce para que estén acordes a los valores que maneja la empresa, y luego se calcula el monto a pagar en base al valor recortado. Estos recortes no se realizan con prestadores con los que la empresa de medicina prepagada tiene negociado de antemano precios fijos sino más bien con prestadores externos. Este mecanismo es utilizado por todas las empresas de medicina prepagada y sirve para evitar estafas, sin embargo, muchas veces los afiliados prefieren acudir a prestadores médicos que no tienen relación con la empresa, con precios muy por encima del mercado y en ocasiones de dudosa reputación.

Reembolso.- Es el pago de los gastos médicos que hace la empresa a los afiliados que han sufrido algún siniestro o enfermedad o a los prestadores médicos que atendieron a un afiliado.

Siniestralidad.- Es el cociente entre el valor total de reembolsos pagados y la facturación Total del cliente.

Valor pagado neto.- Es la suma del valor pagado al afiliado o al prestador que presentó una factura por tratamientos, consultas o medicamentos que haya recibido el afiliado sin tomar en cuenta si tienen o tuvieron cobertura o no según las obligaciones contractuales, es decir, es el valor que la empresa de medicina prepagada cubrió por gastos médicos del afiliado. Según las clausulas del contrato celebrado con el afiliado o la empresa que representa al afiliado se cubre un porcentaje de los gasto (Por lo general la empresa Humana S.A. cubre el 90 % de los gastos).

Valor presentado.- Es la suma del costo total de los tratamientos, consultas o medicamentos que haya recibido el afiliado y que hayan sido presentados a la empresa. No se toma en cuenta si tienen o tuvieron cobertura o no según las obligaciones contractuales. Estos valores se obtienen de las facturas presentadas a la empresa de medicina prepagada por el afiliado o el prestador.

7.4 DICCIONARIO DE VARIABLES

CAMBIO_PLAN_12M.- Variable binaria. Toma el valor de 1 si se cambió el plan del contrato en los últimos 12 meses y 0 en caso contrario.

CAMBIO_PLAN_3M.- Variable binaria. Toma el valor de 1 si se cambió el plan del contrato en los últimos 3 meses y 0 en caso contrario.

CAMBIO_PLAN_6M.- Variable binaria. Toma el valor de 1 si se cambió el plan del contrato en los últimos 6 meses y 0 en caso contrario.

CAMBIO_PLAN_9M.- Variable binaria. Toma el valor de 1 si se cambió el plan del contrato en los últimos 9 meses y 0 en caso contrario.

CANAL.- Variable binaria. Toma el valor de 1 si el canal a través del cual el contrato fue suscrito fue el canal directo, es decir, fue suscrito por un ejecutivo de ventas de Humana S.A., y 0 en caso de haber sido suscrito por otros canales (Bróker o página web).

CENSURA.- Clasifica cada sujeto como censurado o no. Tiene el valor de 1 si el sujeto es no censurado, es decir que, si le ocurrió el evento de interés dentro del periodo de tiempo del estudio, y 0 si no le ocurrió el evento de interés.

CONYUGE.- Variable binaria. Toma el valor 1 si el titular del contrato tiene cónyuge y está incluido en contrato y 0 en caso contrario.

DIAS_FIN_CONTRATO.- Número de días para que termine la versión actual del contrato.

EDAD_TITULAR.- Edad del titular del contrato.

EXCLUSIONES_12M.- Número de exclusiones, es decir, que se excluyó a una persona del contrato, en los últimos 12 meses.

EXCLUSIONES_3M.- Número de exclusiones, es decir, que se excluyó a una persona del contrato, en los últimos 3 meses.

EXCLUSIONES_6M.- Número de exclusiones, es decir, que se excluyó a una persona del contrato, en los últimos 6 meses.

EXCLUSIONES_9M.- Número de exclusiones, es decir, que se excluyó a una persona del contrato, en los últimos 9 meses.

GENERO_TITULAR.- Genero del titular del contrato.

INCLUSIONES_12M.- Número de inclusiones, es decir, que se incluyó a una persona más en el contrato, en los últimos 12 meses.

INCLUSIONES_3M.- Número de inclusiones, es decir, que se incluyó a una persona más en el contrato, en los últimos 3 meses.

INCLUSIONES_6M.- Número de inclusiones, es decir, que se incluyó a una persona más en el contrato, en los últimos 6 meses.

INCLUSIONES_9M.- Número de inclusiones, es decir, que se incluyó a una persona más en el contrato, en los últimos 9 meses.

INCREMENTO_ABS.- Incremento real relativo de precios por concepto de primas (ver anexo 7.3) que se efectuó en el año 2018 (año anterior a l corte de la data).

INCREMENTO_ABSOLUTO_2018.- Incremento presupuestado absoluto de precios por concepto de primas (ver anexo 7.3) que se efectuó en el año 2018 (año anterior a l corte de la data).

INCREMENTO_REL.- Incremento real absoluto de precios por concepto de primas (ver anexo 7.3) que se efectuó en el año 2018 (año anterior a l corte de la data).

INCREMENTO_RELATIVO_2018.- Incremento presupuestado relativo de precios por concepto de primas (ver anexo 7.3) que se efectuó en el año 2018 (año anterior a l corte de la data).

INDIVIDUAL_FAMILIAR.- Variable dicotómica. Toma el valor de 1 cuando el contrato solo tiene un afiliado y 2 si tiene más de uno.

MAX_TIEMPO_OCURRENCIA_PAGO_12M.- Máximo del número días que tomo un reembolsos entre que sobrevino el siniestro y se pagó el reembolso para todos los reembolsos en los últimos 12 meses.

MAX_TIEMPO_OCURRENCIA_PAGO_3M.- Máximo del número días que tomo un reembolsos entre que sobrevino el siniestro y se pagó el reembolso para todos los reembolsos en los últimos 3 meses.

MAX_TIEMPO_OCURRENCIA_PAGO_6M.- Máximo del número días que tomo un reembolsos entre que sobrevino el siniestro y se pagó el reembolso para todos los reembolsos en los últimos 6 meses.

MAX_TIEMPO_OCURRENCIA_PAGO_9M.- Máximo del número días que tomo un reem-

bolsos entre que sobrevino el siniestro y se pagó el reembolso para todos los reembolsos en los últimos 9 meses.

MORA_MAX_12M.- Días de mora máximos en los últimos 12 meses.

MORA_MAX_3M.- Días de mora máximos en los últimos 3 meses.

MORA_MAX_6M.- Días de mora máximos en los últimos 6 meses.

MORA_PROM_12M.- Días de mora en promedio en los últimos 12 meses.

MORA_PROM_3M.- Días de mora en promedio en los últimos 3 meses.

MORA_PROM_6M.- Días de mora en promedio en los últimos 6 meses.

MOVIMIENTOS_12M.- Número de movimientos (inclusiones, exclusiones o cambios de plan) que se realizó en el contrato en los últimos 12 meses.

MOVIMIENTOS_3M.- Número de movimientos (inclusiones, exclusiones o cambios de plan) que se realizó en el contrato en los últimos 3 meses.

MOVIMIENTOS_6M.- Número de movimientos (inclusiones, exclusiones o cambios de plan) que se realizó en el contrato en los últimos 6 meses.

MOVIMIENTOS_9M.- Número de movimientos (inclusiones, exclusiones o cambios de plan) que se realizó en el contrato en los últimos 9 meses.

NUEVO.- Variable binaria. Toma el valor de 1 si el contrato no tiene más de tres meses de afiliación a la compañía y cero en caso contrario.

NUMERO_AFILIADOS.- Número de afiliados del contrato.

NUMERO_AFILIADOS_2AÑOS_O_MENOS.- Número de afiliados menores de 2 años del contrato.

NUMERO_AFILIADOS_5AÑOS_O_MENOS.- Número de afiliados menores de 5 años del contrato.

NUMERO_AFILIADOS_MENORES.- Número de afiliados menores de edad del contrato.

PAGADO_NETO.- Valor pagado neto (ver anexo 7.3) en los últimos 3 años. Hubiese sido deseable tener el valor pagado neto durante toda la relación contractual, sin embargo, la empresa auspiciante limitó los datos a los últimos 3 años.

PAGADO_NETO_12M.- Valor pagado neto (ver anexo 7.3) en los últimos 12 meses.

PAGADO_NETO_12M_PA.- Valor pagado neto (ver anexo 7.3) en los últimos 12 meses únicamente en la modalidad de pago al afiliado (ver anexo 7.3).

PAGADO_NETO_3M.- Valor pagado neto (ver anexo 7.3) en los últimos 3 meses.

PAGADO_NETO_3M_PA.- Valor pagado neto (ver anexo 7.3) en los últimos 3 meses únicamente en la modalidad de pago al afiliado (ver anexo 7.3).

PAGADO_NETO_6M.- Valor pagado neto (ver anexo 7.3) en los últimos 6 meses.

PAGADO_NETO_6M_PA.- Valor pagado neto (ver anexo 7.3) en los últimos 6 meses únicamente en la modalidad de pago al afiliado (ver anexo 7.3).

PAGADO_NETO_9M.- Valor pagado neto (ver anexo 7.3) en los últimos 9 meses.

PAGADO_NETO_9M_PA.- Valor pagado neto (ver anexo 7.3) en los últimos 9 meses únicamente en la modalidad de pago al afiliado (ver anexo 7.3).

PAGADO_NETO_CPC_12M.- Valor pagado neto (ver anexo 7.3) en los últimos 12 meses únicamente por enfermedades crónicas o potencialmente crónicas (ver anexo 7.3).

PAGADO_NETO_CRONICO_12M.- Valor pagado neto (ver anexo 7.3) en los últimos 12 meses únicamente por enfermedades crónicas (ver anexo 7.3).

PAGOS_DEBITO_BANCARIO_12M.- Monto total pagado mediante débito bancario por el afiliado en los últimos 12 meses por concepto de primas (ver anexo 7.3).

PAGOS_DEBITO_BANCARIO_3M.- Monto total pagado mediante débito bancario por el afiliado en los últimos 3 meses por concepto de primas (ver anexo 7.3).

PAGOS_DEBITO_BANCARIO_6M.- Monto total pagado mediante débito bancario por el afiliado en los últimos 6 meses por concepto de primas (ver anexo 7.3).

PAGOS_EFECTIVO_12M.- Monto total pagado en efectivo por el afiliado en los últimos 12 meses por concepto de primas (ver anexo 7.3).

PAGOS_EFECTIVO_3M.- Monto total pagado en efectivo por el afiliado en los últimos 3 meses por concepto de primas (ver anexo 7.3).

PAGOS_EFECTIVO_6M.- Monto total pagado en efectivo por el afiliado en los últimos 6 meses por concepto de primas (ver anexo 7.3).

PAGOS_TARJETA_CREDITO_12M.- Monto total pagado mediante tarjeta de crédito por el

afiliado en los últimos 12 meses por concepto de primas (ver anexo 7.3).

PAGOS_TARJETA_CREDITO_3M.- Monto total pagado mediante tarjeta de crédito por el afiliado en los últimos 3 meses por concepto de primas (ver anexo 7.3).

PAGOS_TARJETA_CREDITO_6M.- Monto total pagado mediante tarjeta de crédito por el afiliado en los últimos 6 meses por concepto de primas (ver anexo 7.3).

PORCENTAJE_COBERTURA.- Es el cociente entre las variables PAGADO_NETO y PRESENTADO. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_12M.- Es el cociente entre las variables PAGADO_NETO_12M y PRESENTADO_12M. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_12M_PA.- Es el cociente entre las variables PAGADO_NETO_12M_PA y PRESENTADO_12M_PA. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_3M.- Es el cociente entre las variables PAGADO_NETO_3M y PRESENTADO_3M. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_3M_PA.- Es el cociente entre las variables PAGADO_NETO_3M_PA y PRESENTADO_3M_PA. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_6M.- Es el cociente entre las variables PAGADO_NETO_6M y PRESENTADO_6M. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_6M_PA.- Es el cociente entre las variables PAGADO_NETO_6M_PA y PRESENTADO_6M_PA. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_9M.- Es el cociente entre las variables PAGADO_NETO_9M y PRESENTADO_9M. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_9M_PA.- Es el cociente entre las variables PAGADO_NETO_9M_PA y PRESENTADO_9M_PA. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_CPC_12M.- Es el cociente entre las variables PAGADO_NETO_CPC_12M y PRESENTADO_CPC_12M. Tiene valores ausentes cuando el divisor es 0.

PORCENTAJE_COBERTURA_CRONICO_12M.- Es el cociente entre las variables PAGADO_NETO_CRONICO_12M y PRESENTADO_CRONICO_12M. Tiene valores ausentes cuando el divisor es 0.

PRESENTADO.- Valor presentado (ver anexo 7.3) en los últimos 3 años. Hubiese sido

deseable tener el valor presentado durante toda la relación contractual, sin embargo, la empresa auspiciante limitó los datos a los últimos 3 años.

PRESENTADO_12M.- Valor presentado (ver anexo 7.3) en los últimos 12 meses.

PRESENTADO_12M_PA.- Valor presentado (ver anexo 7.3) en los últimos 12 meses únicamente en la modalidad de pago al afiliado (ver anexo 7.3).

PRESENTADO_3M.- Valor presentado (ver anexo 7.3) en los últimos 3 meses.

PRESENTADO_3M_PA.- Valor presentado (ver anexo 7.3) en los últimos 3 meses únicamente en la modalidad de pago al afiliado (ver anexo 7.3).

PRESENTADO_6M.- Valor presentado (ver anexo 7.3) en los últimos 6 meses.

PRESENTADO_6M_PA.- Valor presentado (ver anexo 7.3) en los últimos 6 meses únicamente en la modalidad de pago al afiliado (ver anexo 7.3).

PRESENTADO_9M.- Valor presentado (ver anexo 7.3) en los últimos 9 meses.

PRESENTADO_9M_PA.- Valor presentado (ver anexo 7.3) en los últimos 9 meses únicamente en la modalidad de pago al afiliado (ver anexo 7.3).

PRESENTADO_CPC_12M.- Valor presentado (ver anexo 7.3) en los últimos 12 meses únicamente por enfermedades crónicas o potencialmente crónicas (ver anexo 7.3).

PRESENTADO_CRONICO_12M.- Valor presentado (ver anexo 7.3) en los últimos 12 meses únicamente por enfermedades crónicas (ver anexo 7.3).

PRIMA_PE.- Prima por expuesto, es decir, la prima promedio por afiliado.

PRODUCTO_MH.- Variable binaria. Toma el valor 1 si el plan del contrato es un producto Metrohumana y 0 si es Practihumana (Solo existen estos dos tipos de producto y se trata de una clasificación interna de la empresa).

RECORTE_12M.- Valor por concepto de recorte (ver anexo 7.3) en los últimos 12 meses.

RECORTE_3M.- Valor por concepto de recorte (ver anexo 7.3) en los últimos 3 meses.

RECORTE_6M.- Valor por concepto de recorte (ver anexo 7.3) en los últimos 6 meses.

RECORTE_9M.- Valor por concepto de recorte (ver anexo 7.3) en los últimos 9 meses.

REEMBOLSOS_CERO_PAGO_12M.- Número de reembolsos en los que el pagado al afiliado es de cero en los últimos 12 meses. Se dan cuando el valor pagado neto no supera el

deducible (ver anexo 7.3).

REEMBOLSOS_CERO_PAGO_3M.- Número de reembolsos en los que el pagado al afiliado es de cero en los últimos 3 meses. Se dan cuando el valor pagado neto no supera el deducible (ver anexo 7.3).

REEMBOLSOS_CERO_PAGO_6M.- Número de reembolsos en los que el pagado al afiliado es de cero en los últimos 6 meses. Se dan cuando el valor pagado neto no supera el deducible (ver anexo 7.3).

REEMBOLSOS_CERO_PAGO_9M.- Número de reembolsos en los que el pagado al afiliado es de cero en los últimos 9 meses. Se dan cuando el valor pagado neto no supera el deducible (ver anexo 7.3).

SINIESTRALIDAD_12M.- Cociente entre el valor reembolsado y el valor facturado en los últimos 12 meses.

SINIESTRALIDAD_6M.- Cociente entre el valor reembolsado y el valor facturado en los últimos 6 meses.

TIEMPO.- Instante en el que le ocurrió el evento de interés. En nuestro estudio, si al sujeto no le ocurrió el evento de interés, el tiempo es el máximo del estudio (En otro tipo de estudios no tiene por qué ocurrir esto).

TIEMPO_AFILIACION.- Tiempo en días desde el inicio del contrato hasta la fecha de corte.

TIEMPO_SIN_USO.- Número días transcurridos desde el últimos uso del servicio de alguno de los afiliados del contrato. Si nunca se ha hecho uso del servicio esta variable es igual al tiempo de afiliación.

TITULAR_SIN_BENEFICIOS.- Variable binaria. Toma el valor 1 si se trata de un titular sin beneficios, es decir, que titular NO está incluido como afiliado en el contrato y 0 en caso contrario.

TOPE_PLAN.- Tope de cobertura del plan del contrato.

USO.- Variable binaria. Toma el valor de 1 si algún afiliado del contrato ha hecho uso del servicio en los últimos 3 años y 0 en caso contrario.

USO_12M.- Variable binaria. Toma el valor de 1 si algún afiliado del contrato ha hecho uso del servicio en los últimos 12 meses y 0 en caso contrario.

USO_6M.- Variable binaria. Toma el valor de 1 si algún afiliado del contrato ha hecho uso del servicio en los últimos 6 meses y 0 en caso contrario.

USO_AMB_EMR_HSD.- Variable binaria. Toma el valor de 1 si algún afiliado del contrato ha hecho uso del servicio por una atención ambulatoria o Emergencia (ver anexo 7.3) en los últimos 3 años y 0 en caso contrario.

USO_HOSPITALARIO.- Variable binaria. Toma el valor de 1 si algún afiliado del contrato ha hecho uso del servicio por una atención hospitalaria (ver anexo 7.3) en los últimos 3 años y 0 en caso contrario.

7.5 ESTIMADOR DE KAPLAN Y MEIER Y PRUEBA LOG-RANK PARA VARIABLES EN LA POBLACIÓN V1

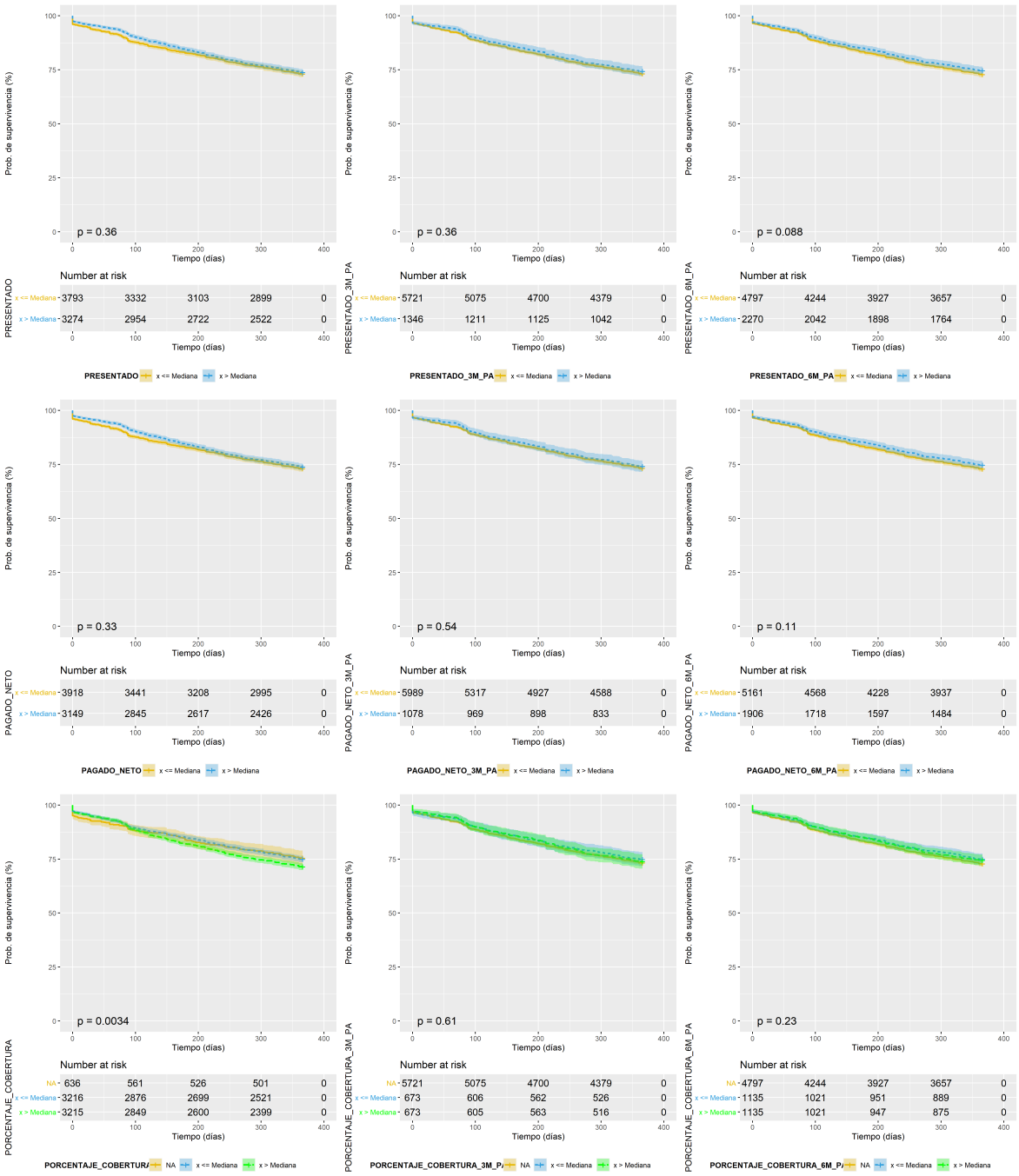


Figura 7.1: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 1

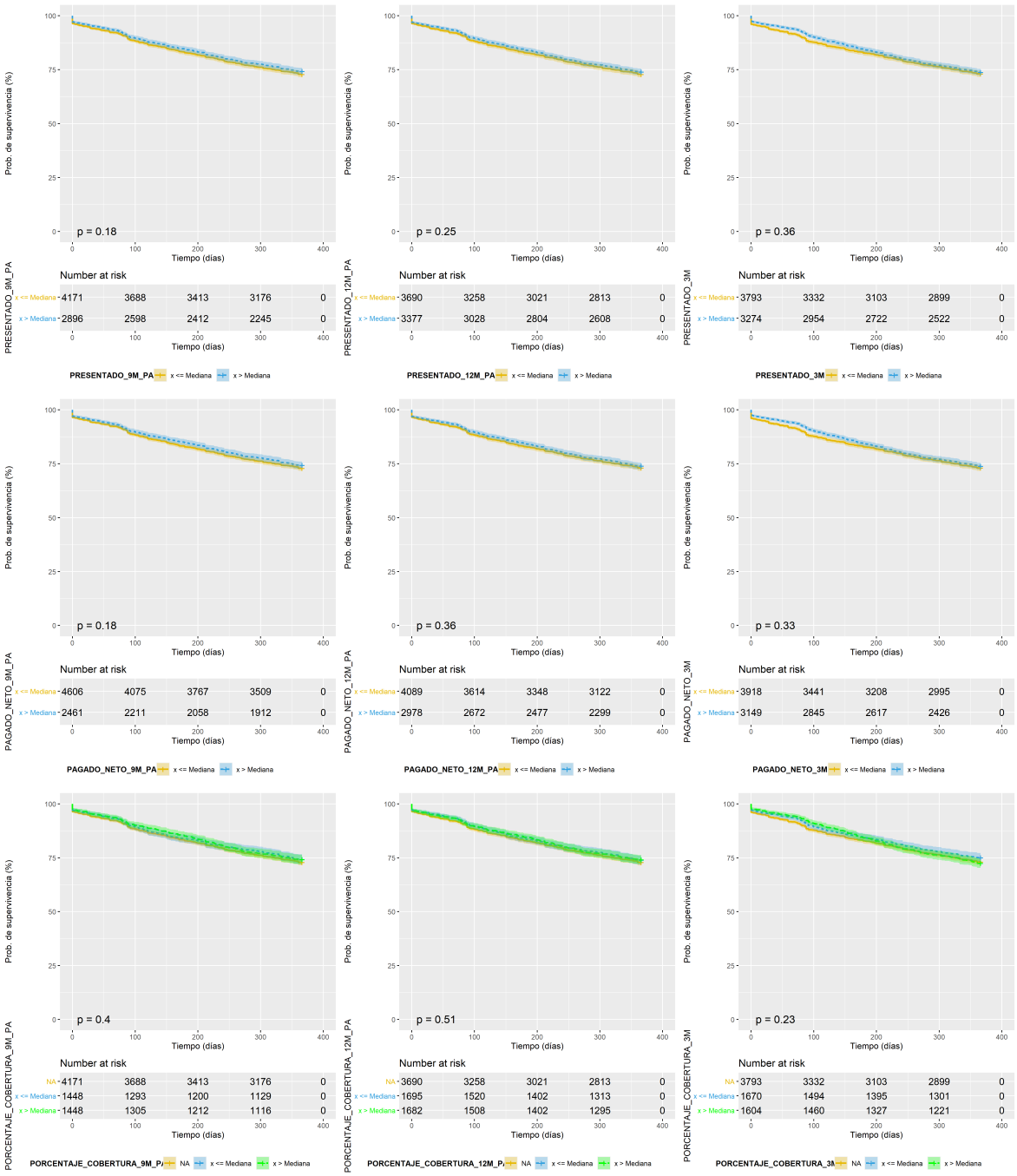


Figura 7.2: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 2

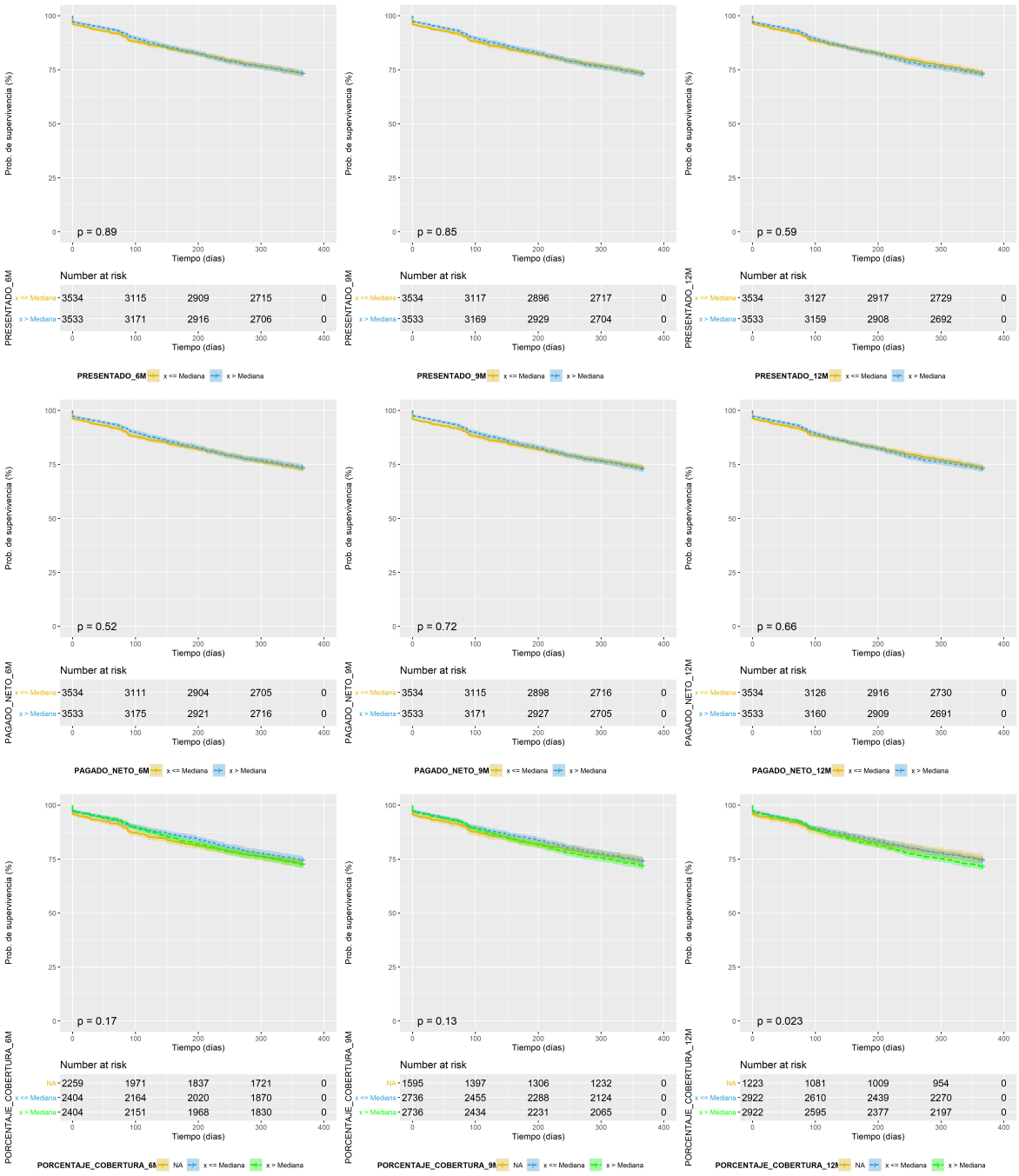


Figura 7.3: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 3

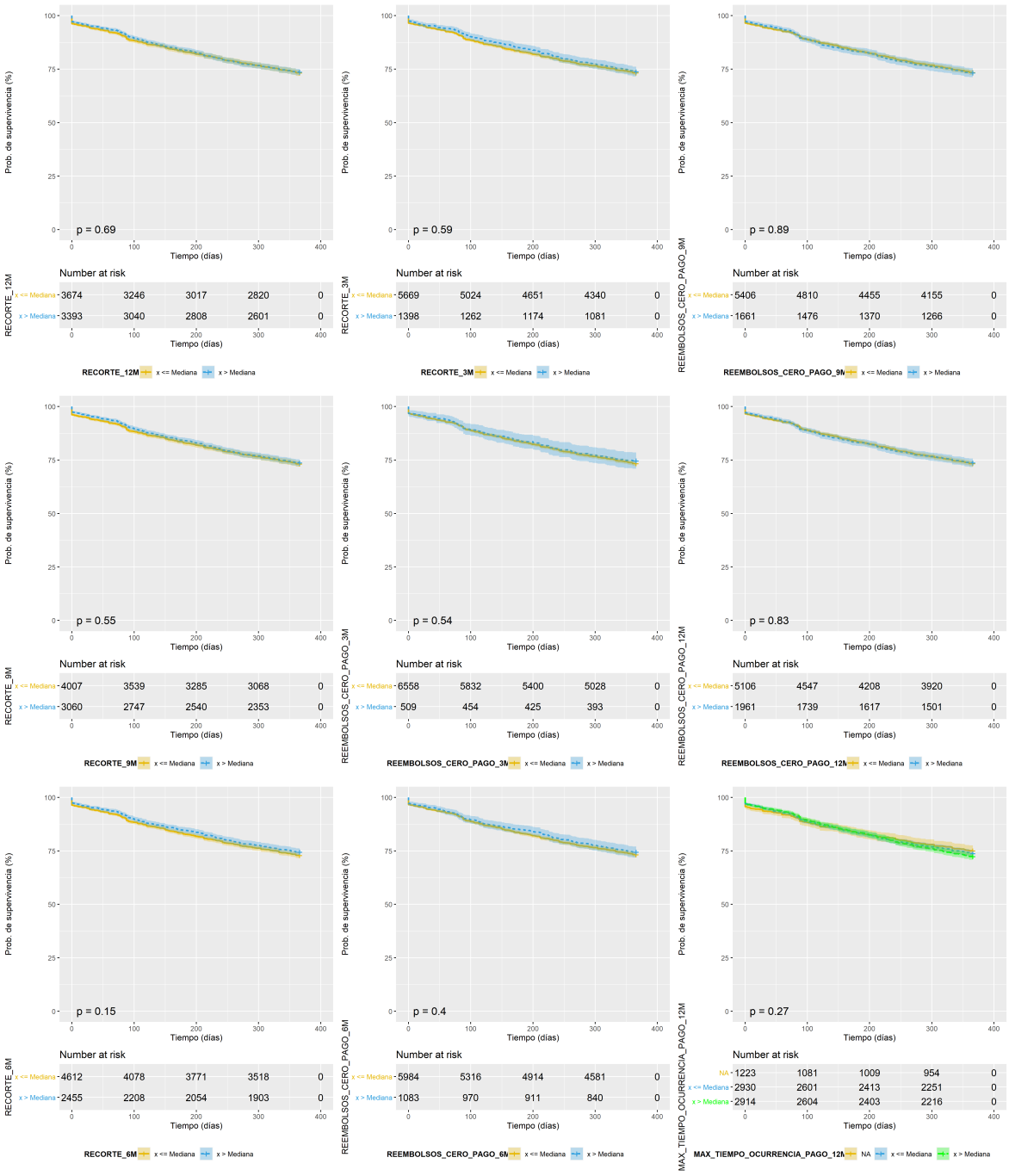


Figura 7.4: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 4

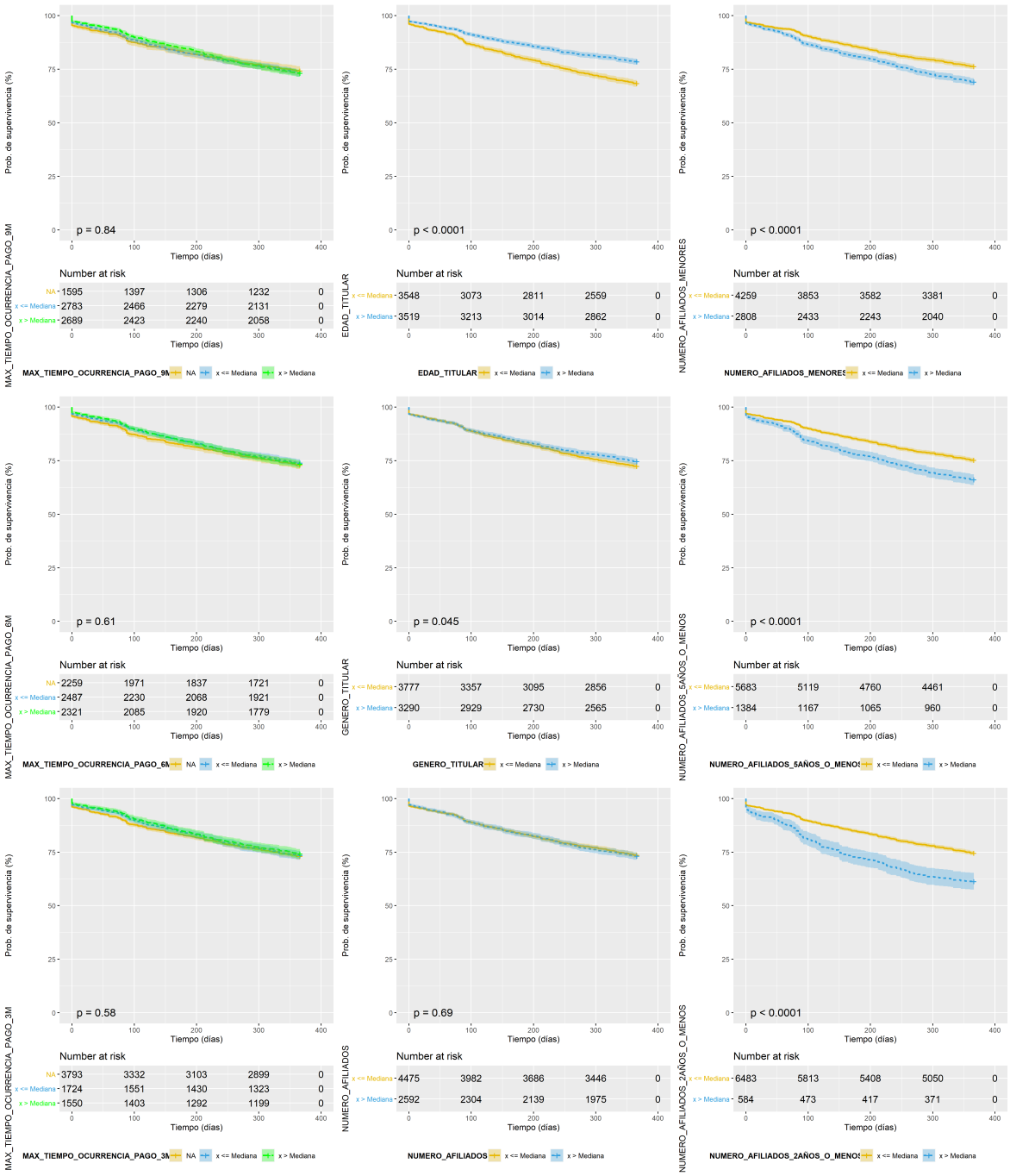


Figura 7.5: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 5

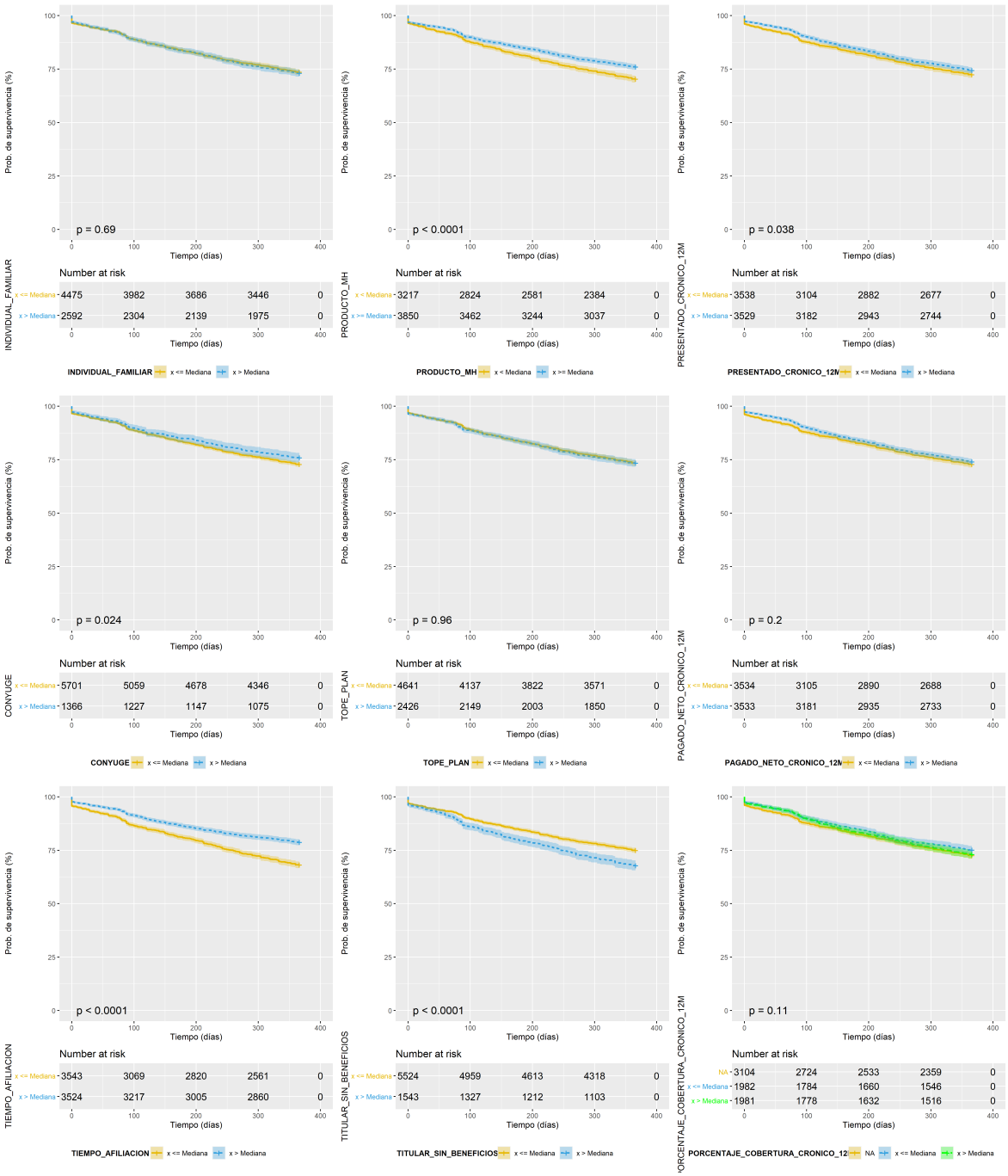


Figura 7.6: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 6

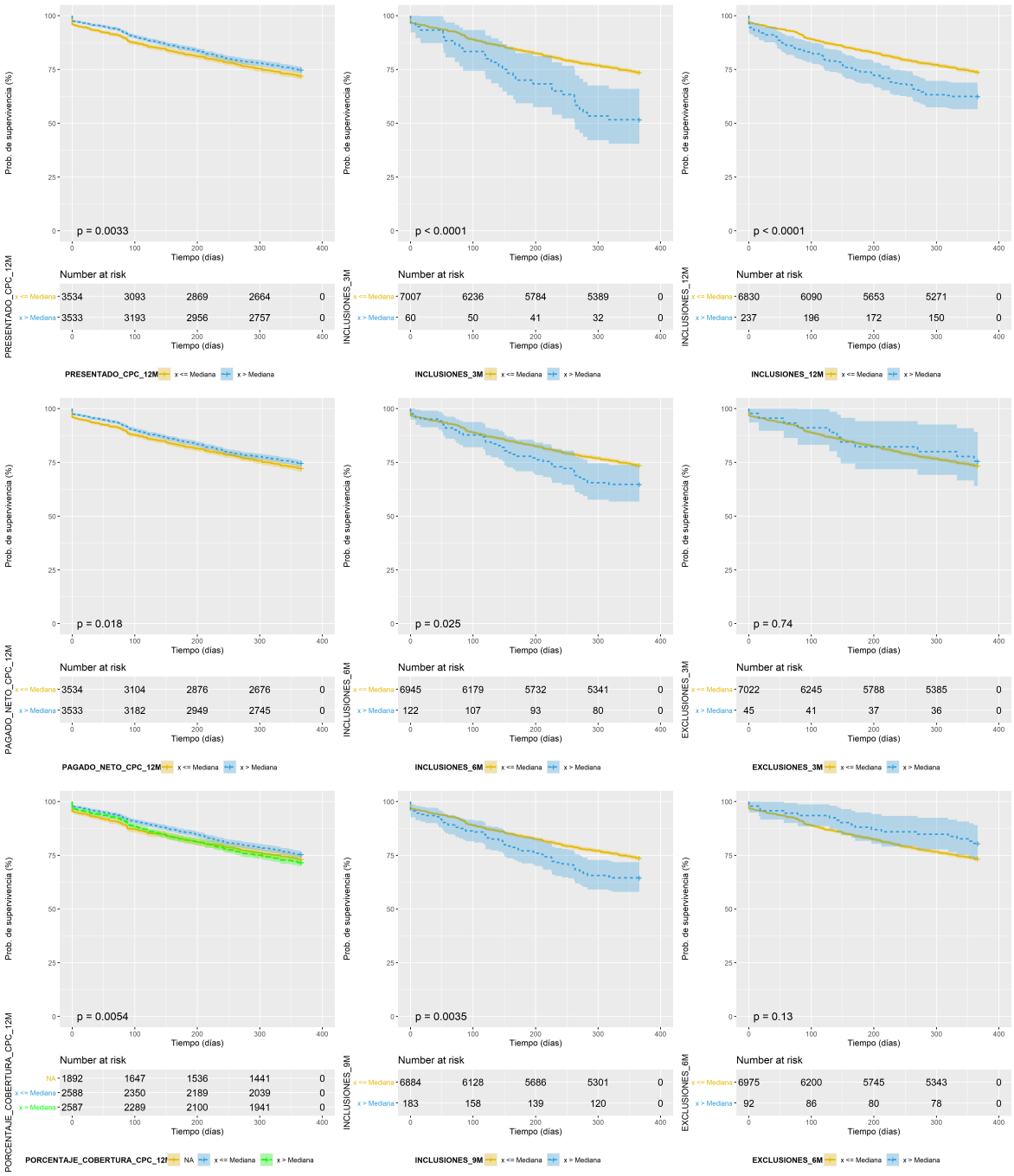


Figura 7.7: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 7

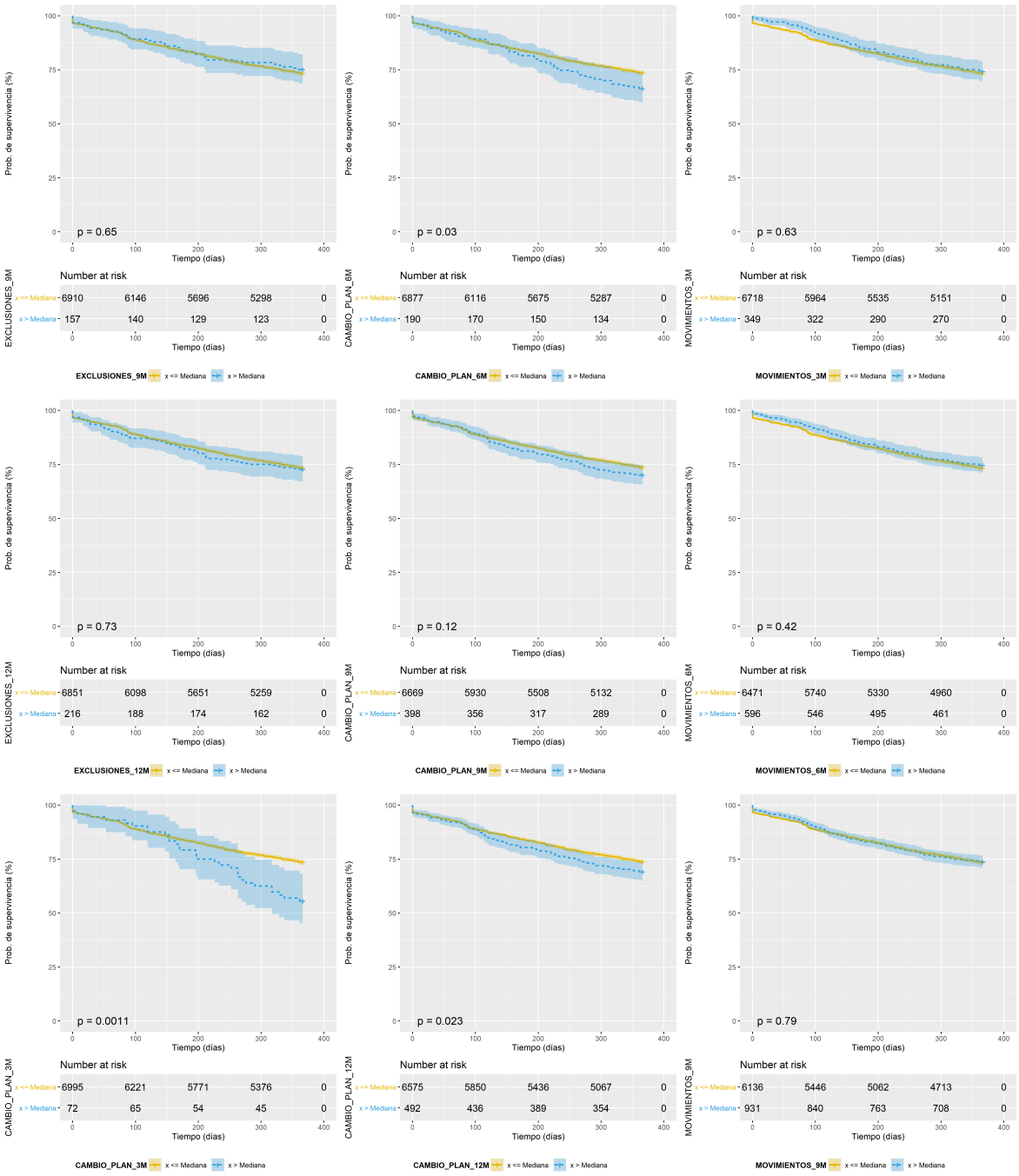


Figura 7.8: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 8

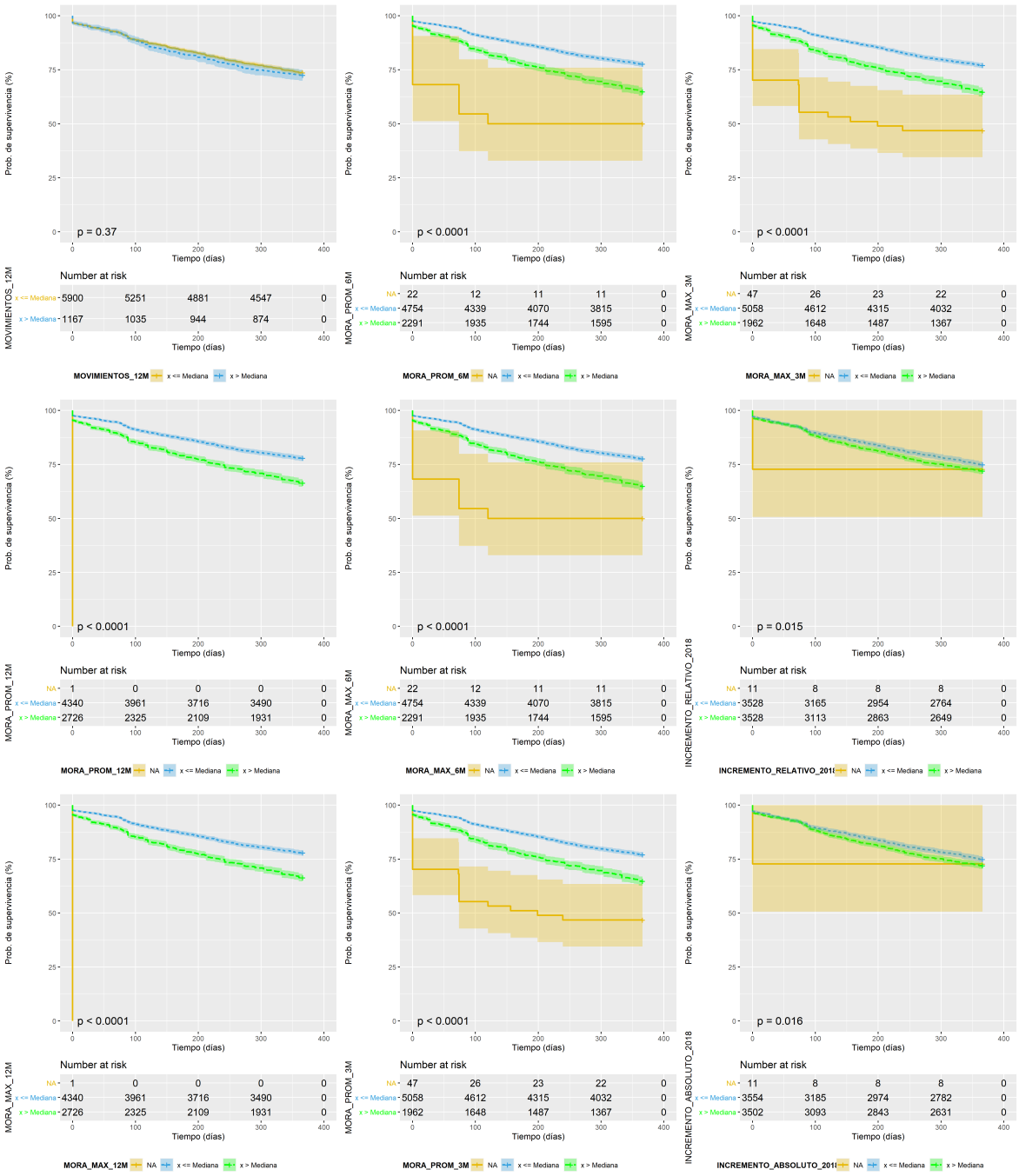


Figura 7.9: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 9

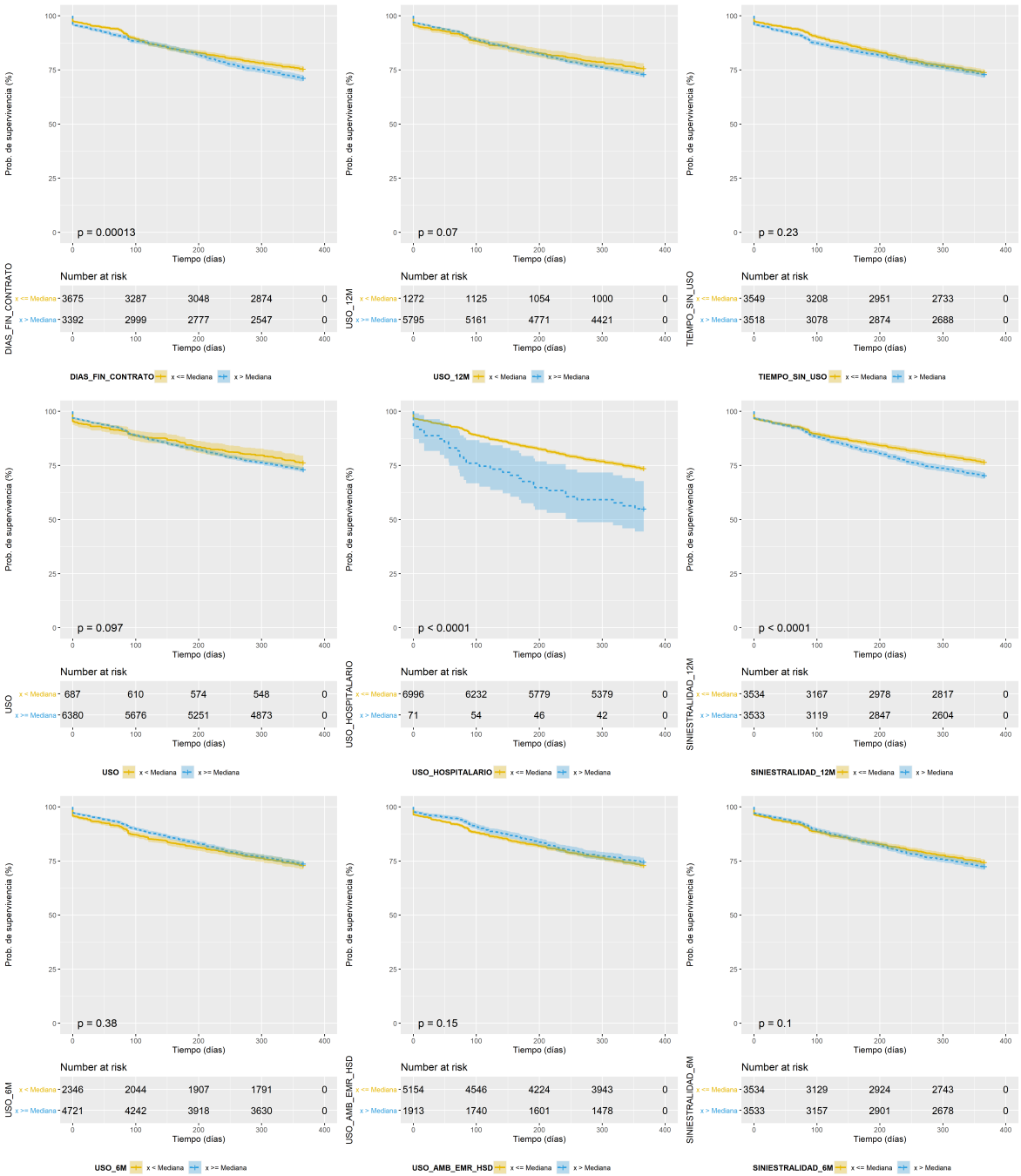


Figura 7.10: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 10

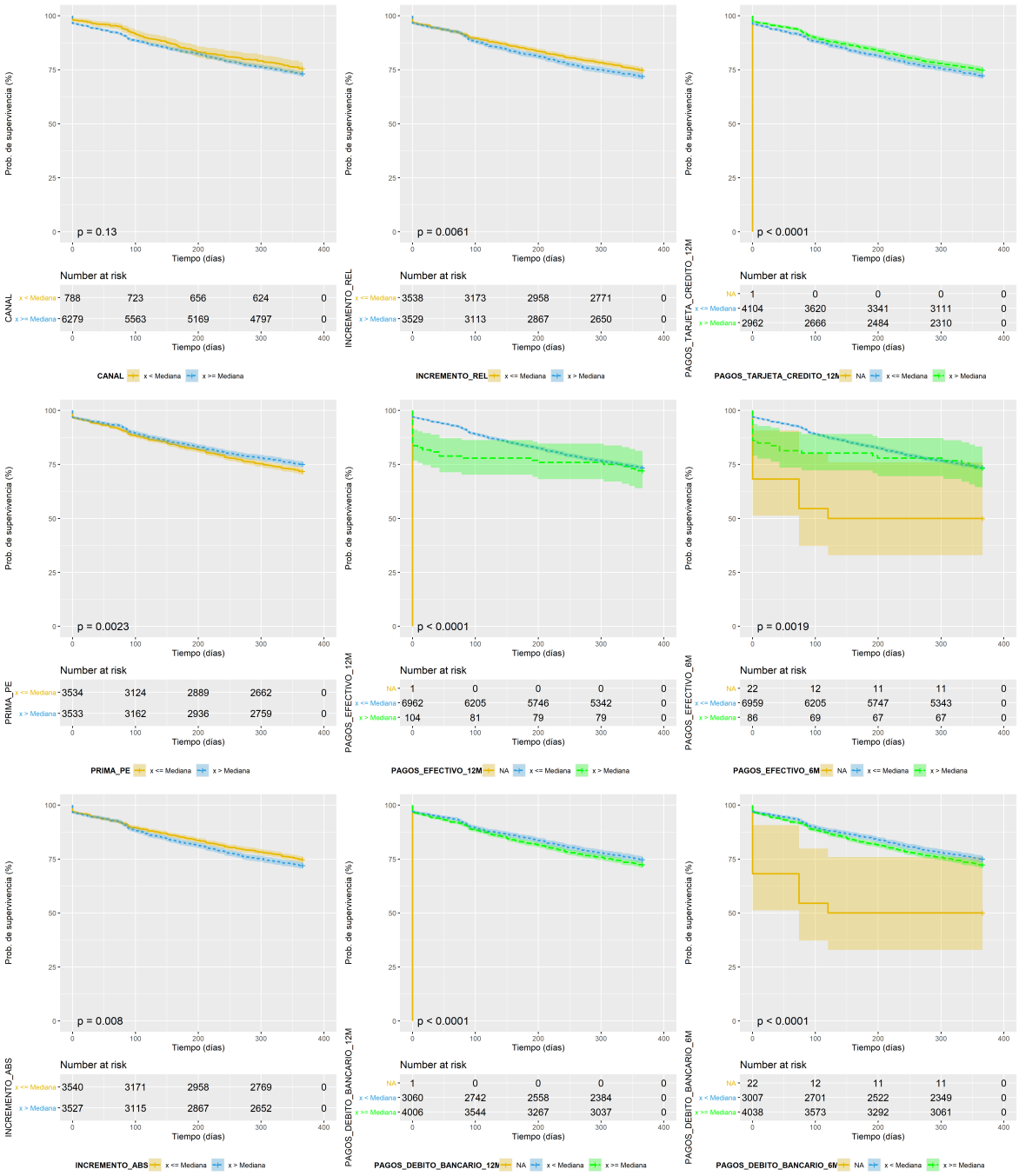


Figura 7.11: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 11

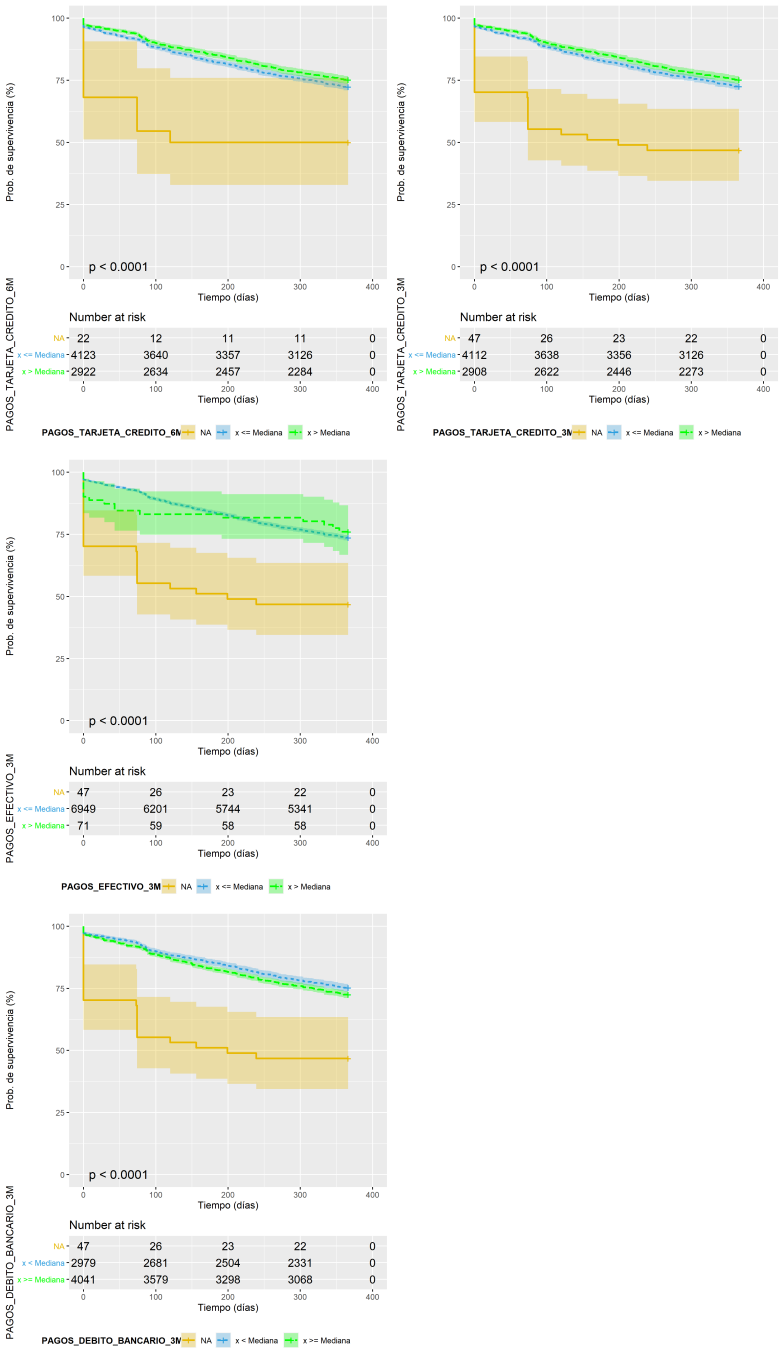


Figura 7.12: Población V1 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 12

7.6 ESTIMADOR DE KAPLAN Y MEIER Y PRUEBA LOG-RANK PARA VARIABLES EN LA POBLACIÓN V2

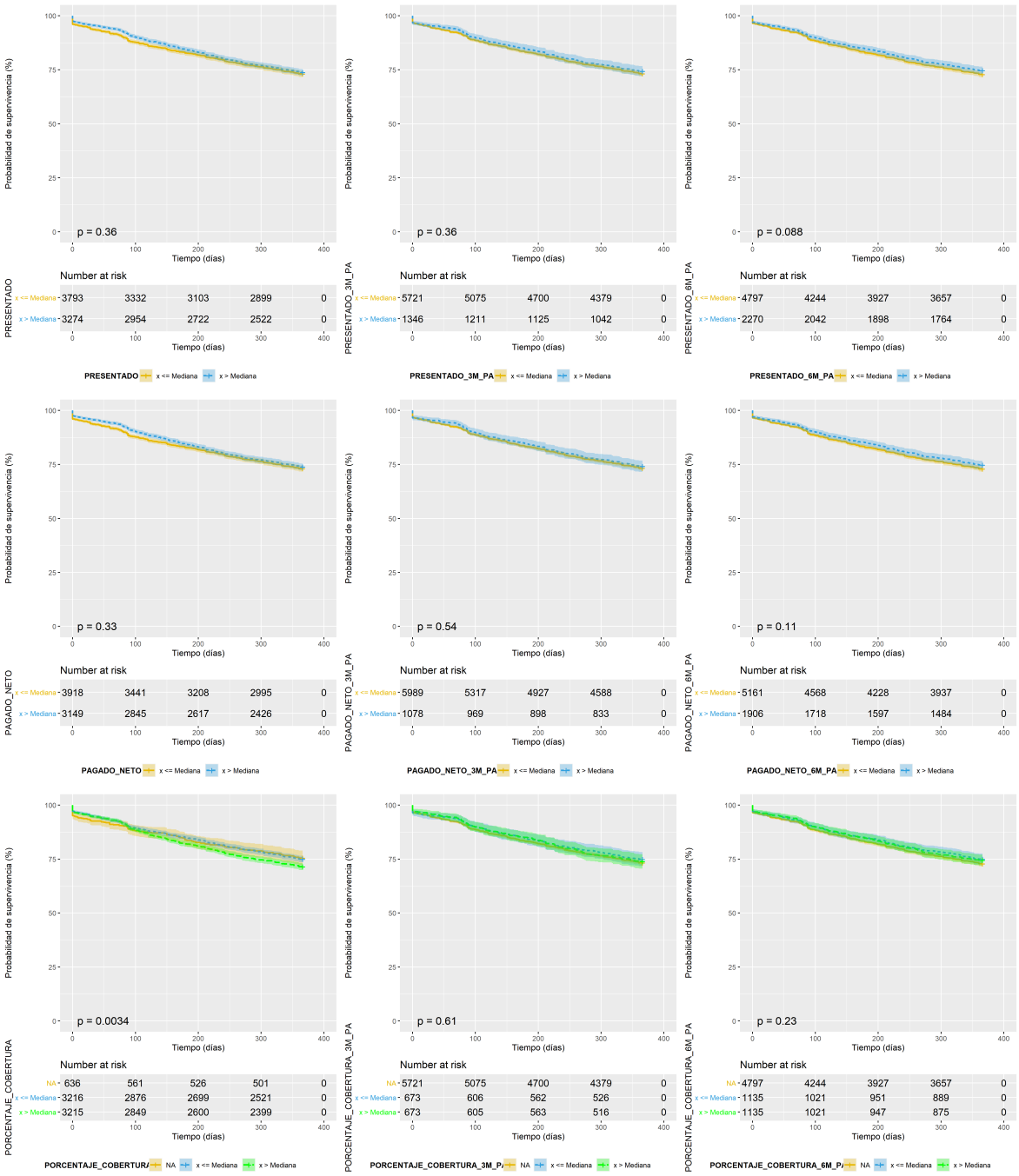


Figura 7.13: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 1

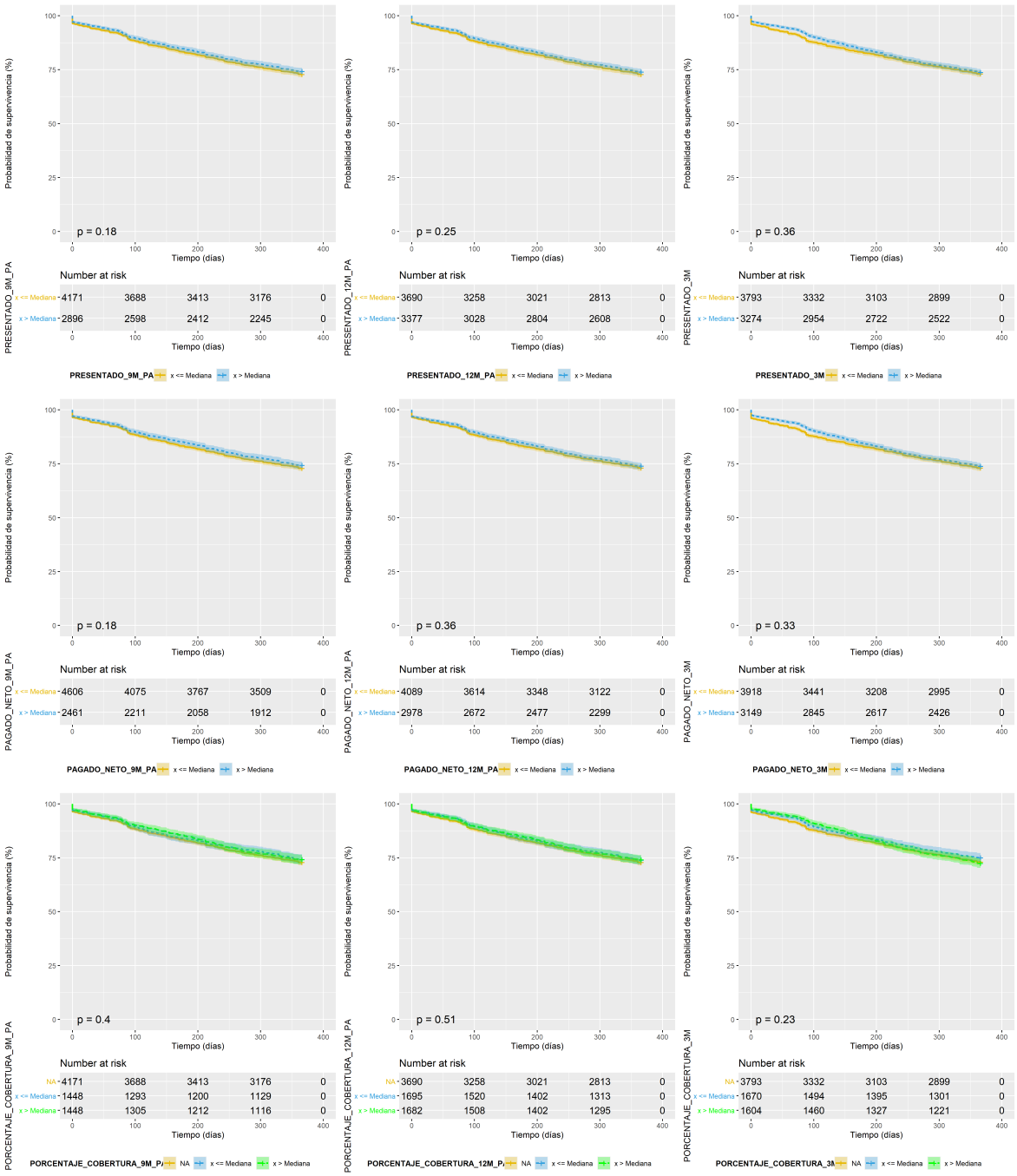


Figura 7.14: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 2

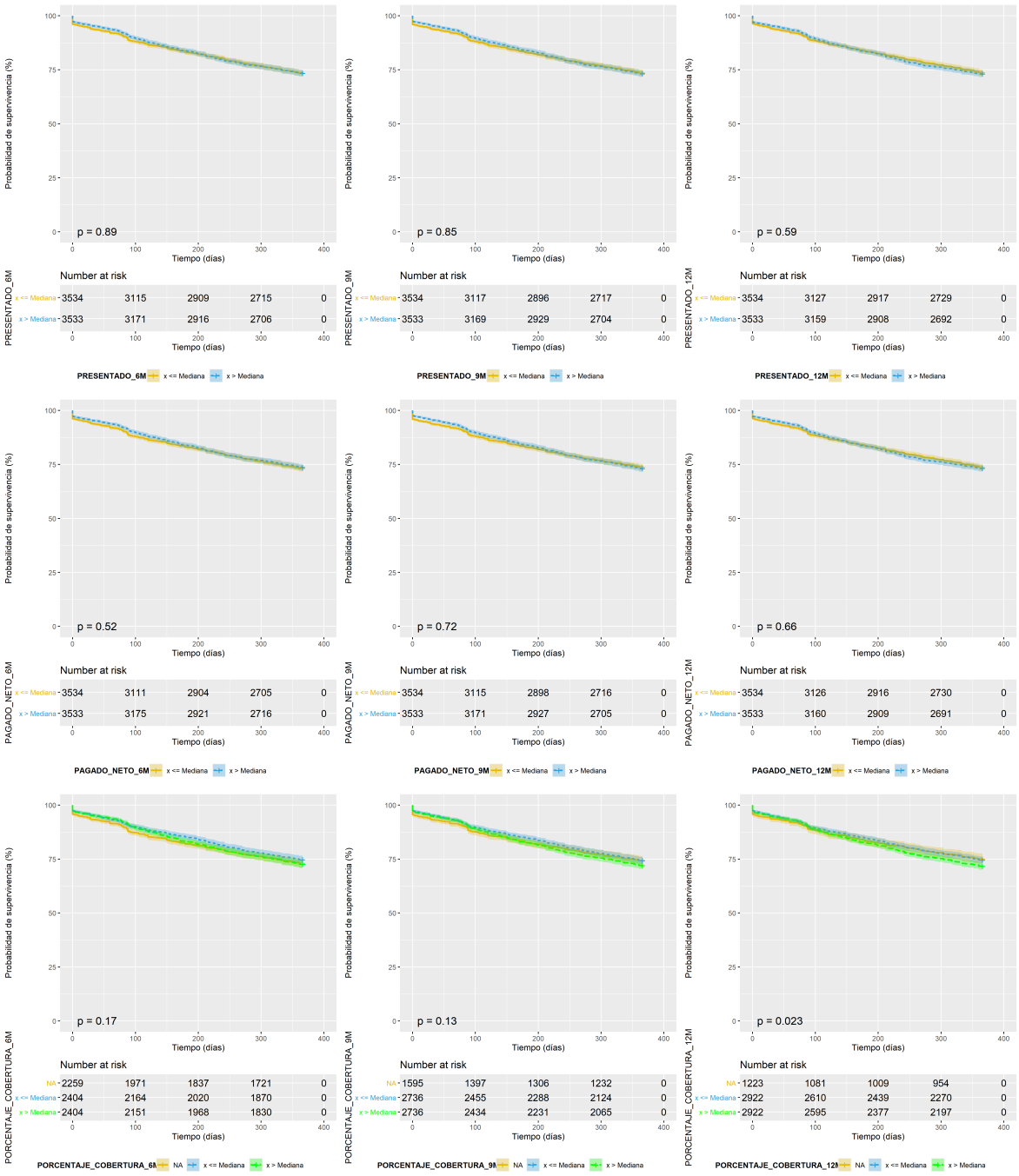


Figura 7.15: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 3

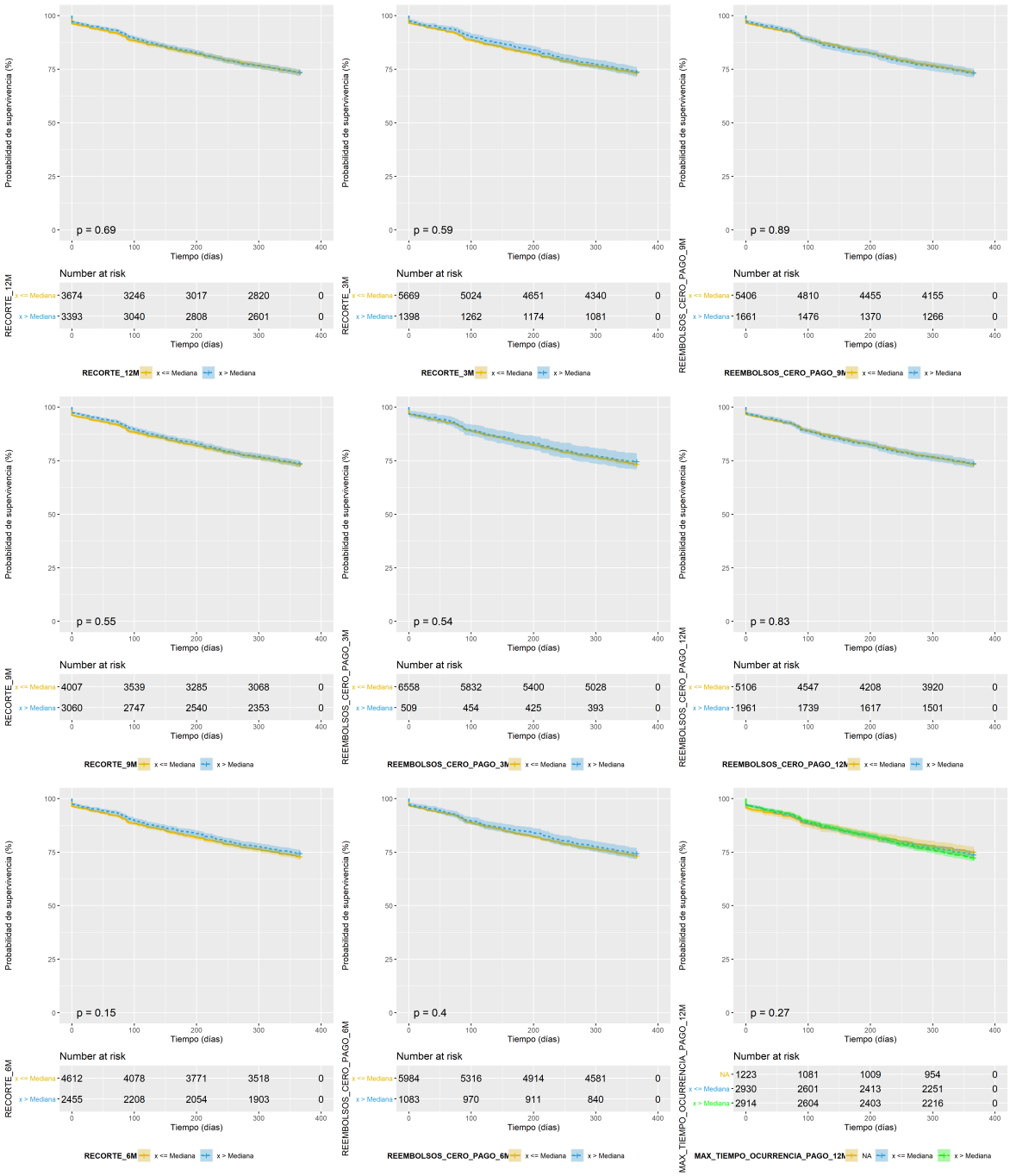


Figura 7.16: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 4

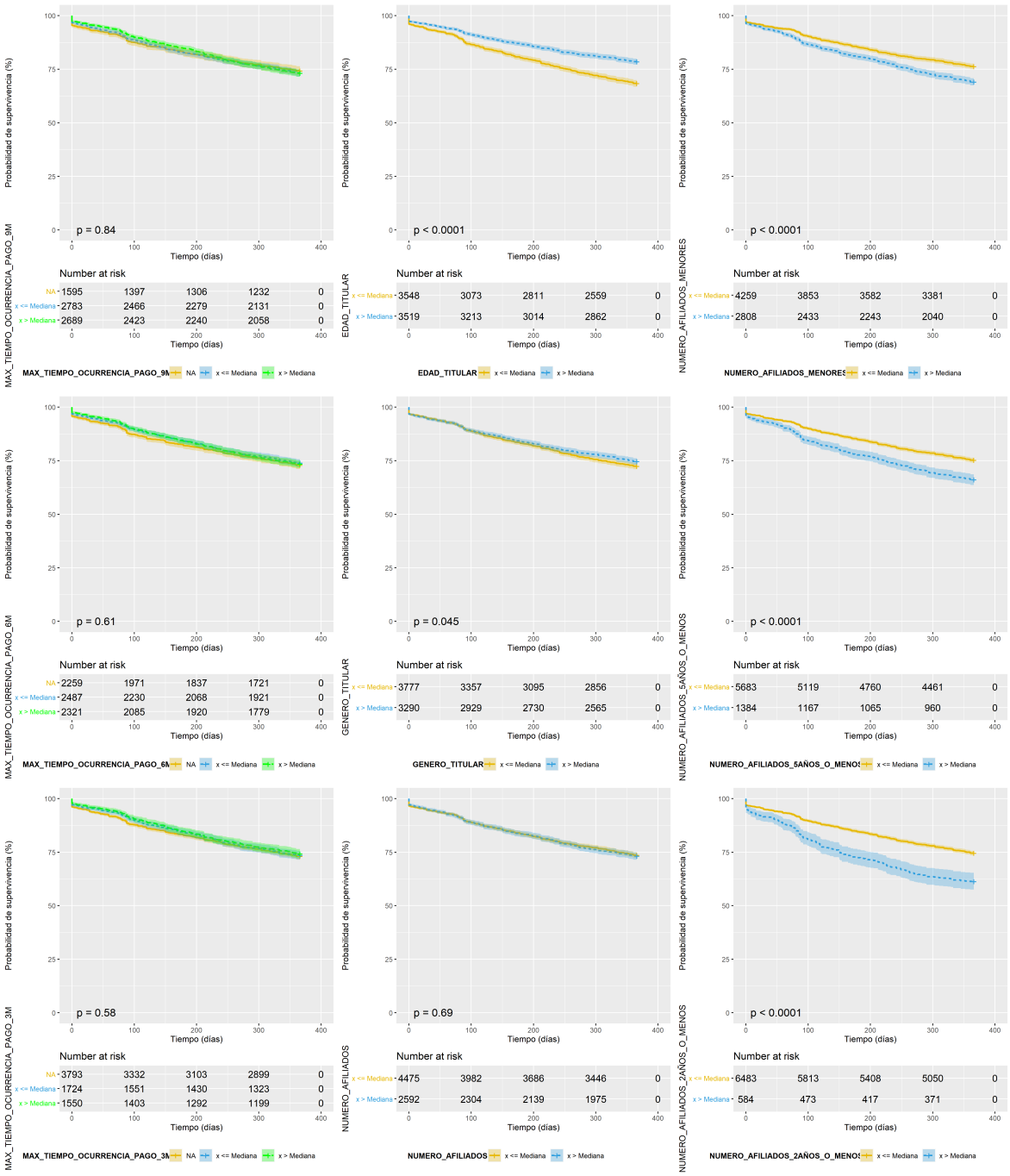


Figura 7.17: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 5

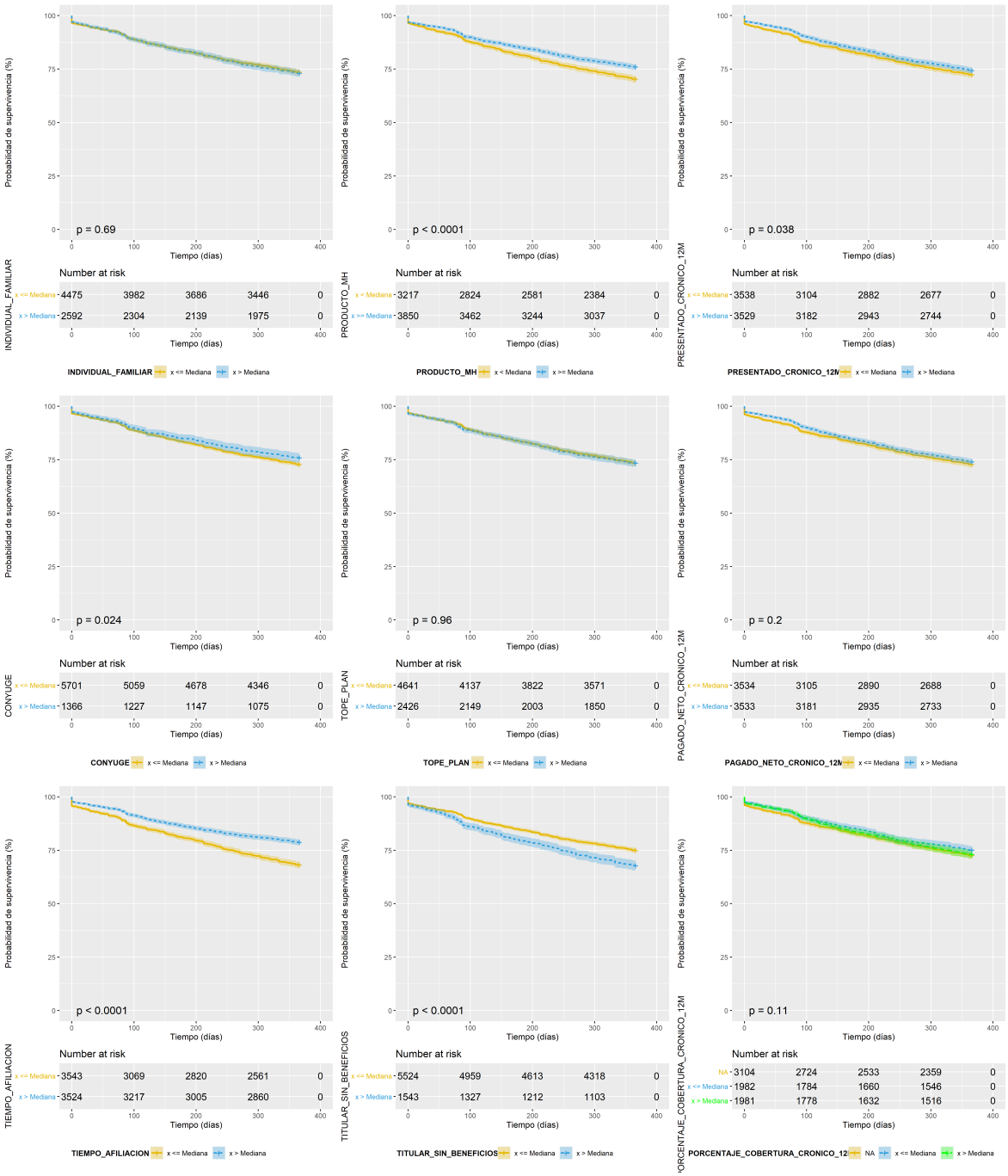


Figura 7.18: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 6

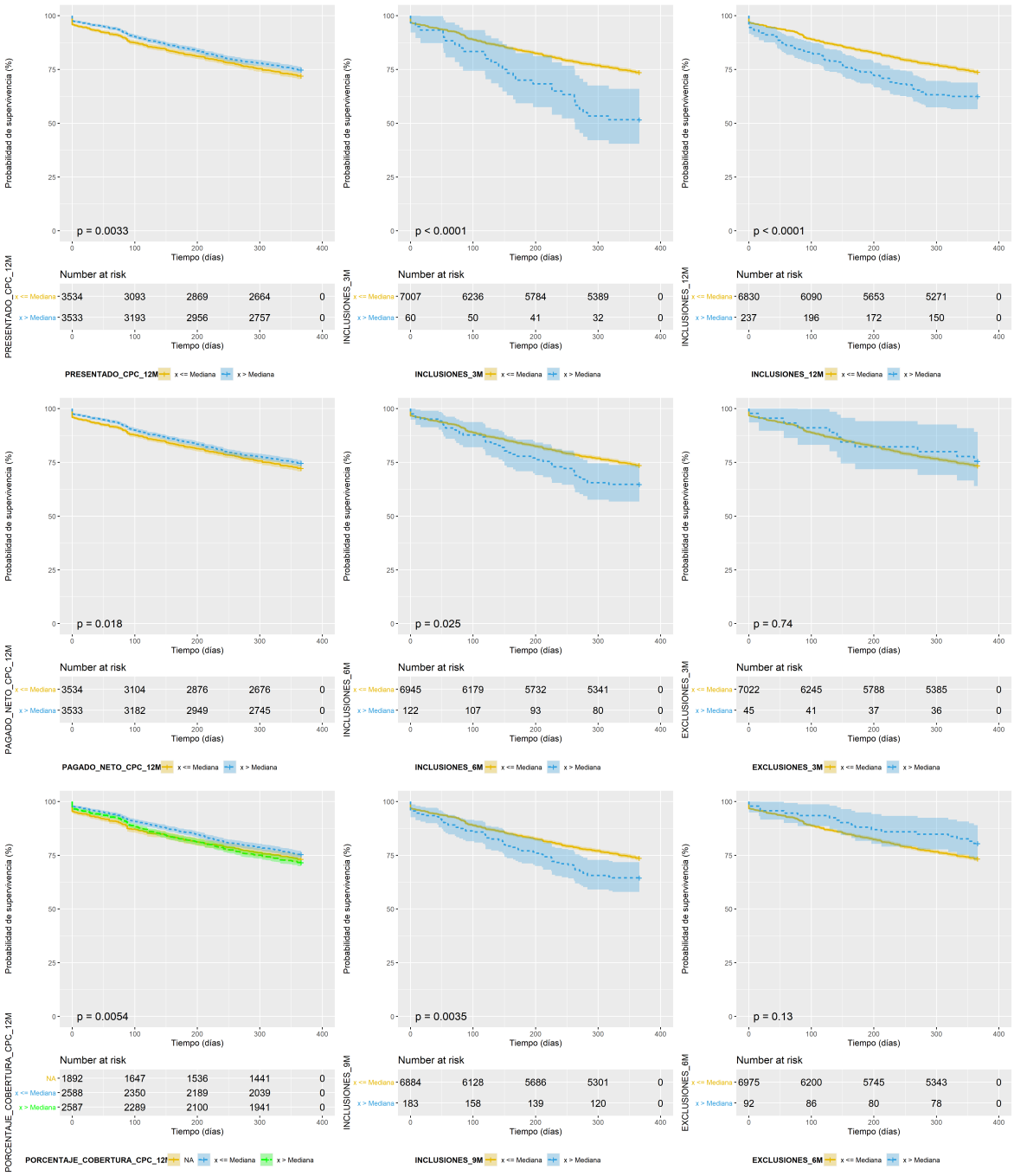


Figura 7.19: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 7

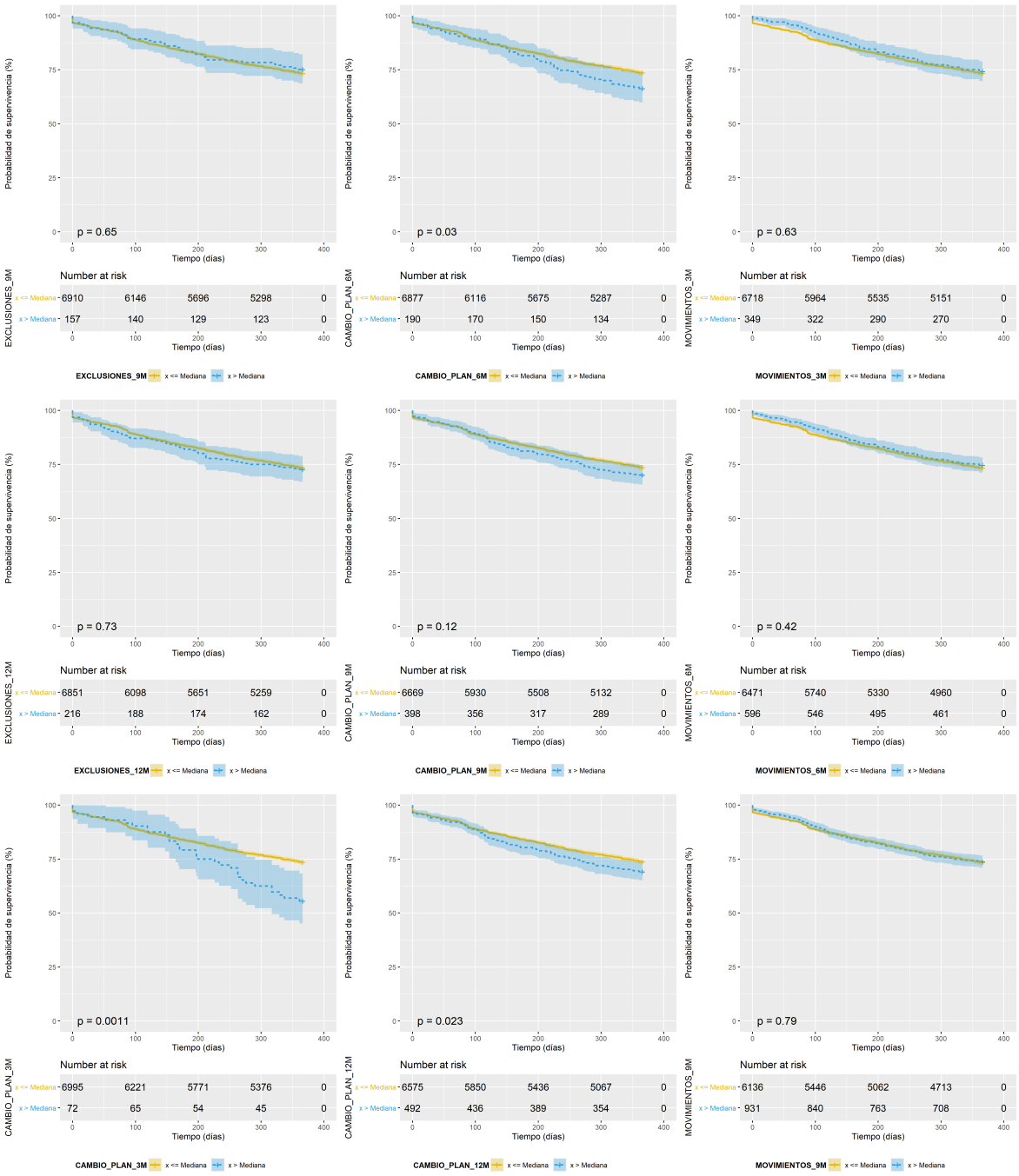


Figura 7.20: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 8

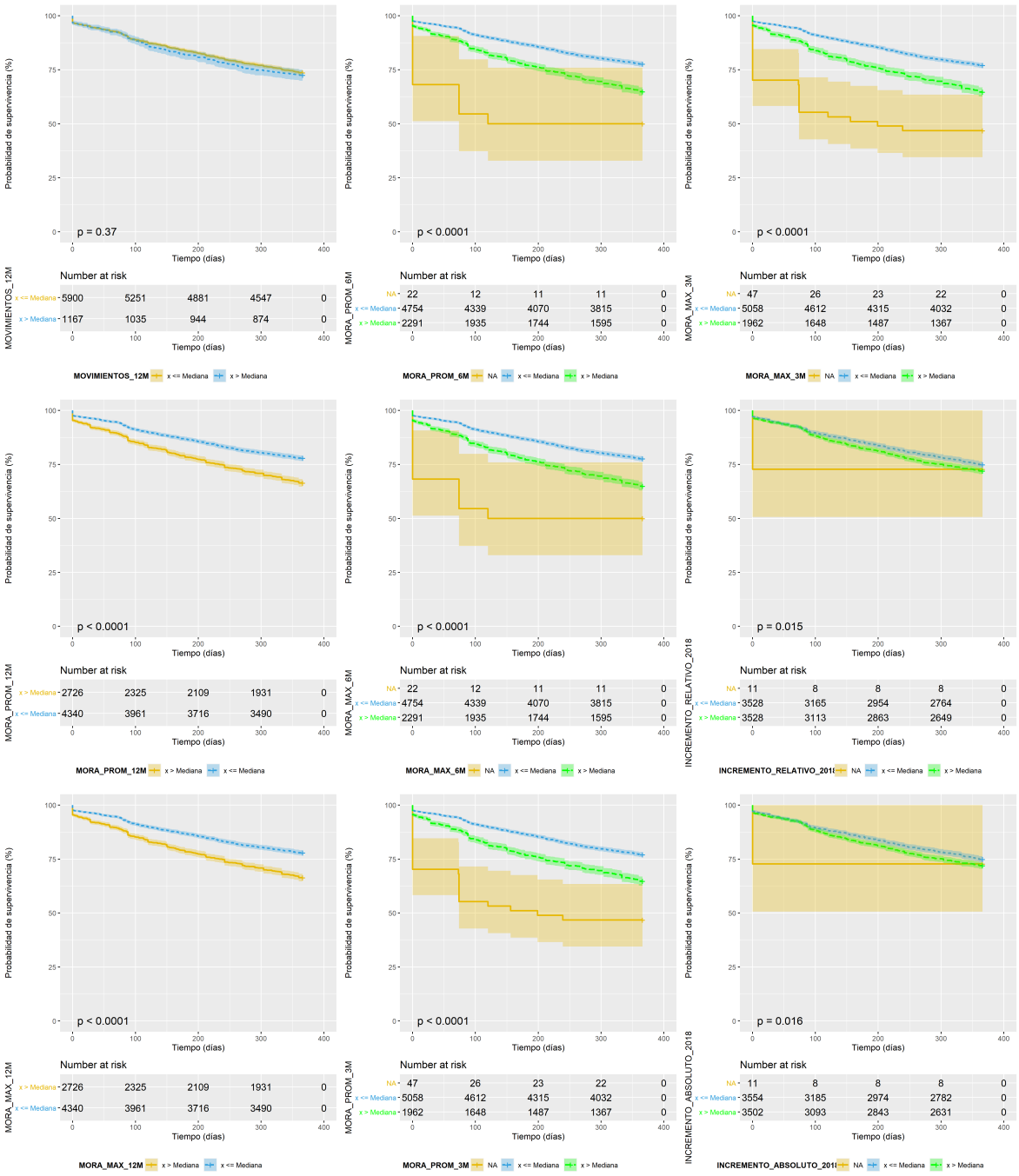


Figura 7.21: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 9

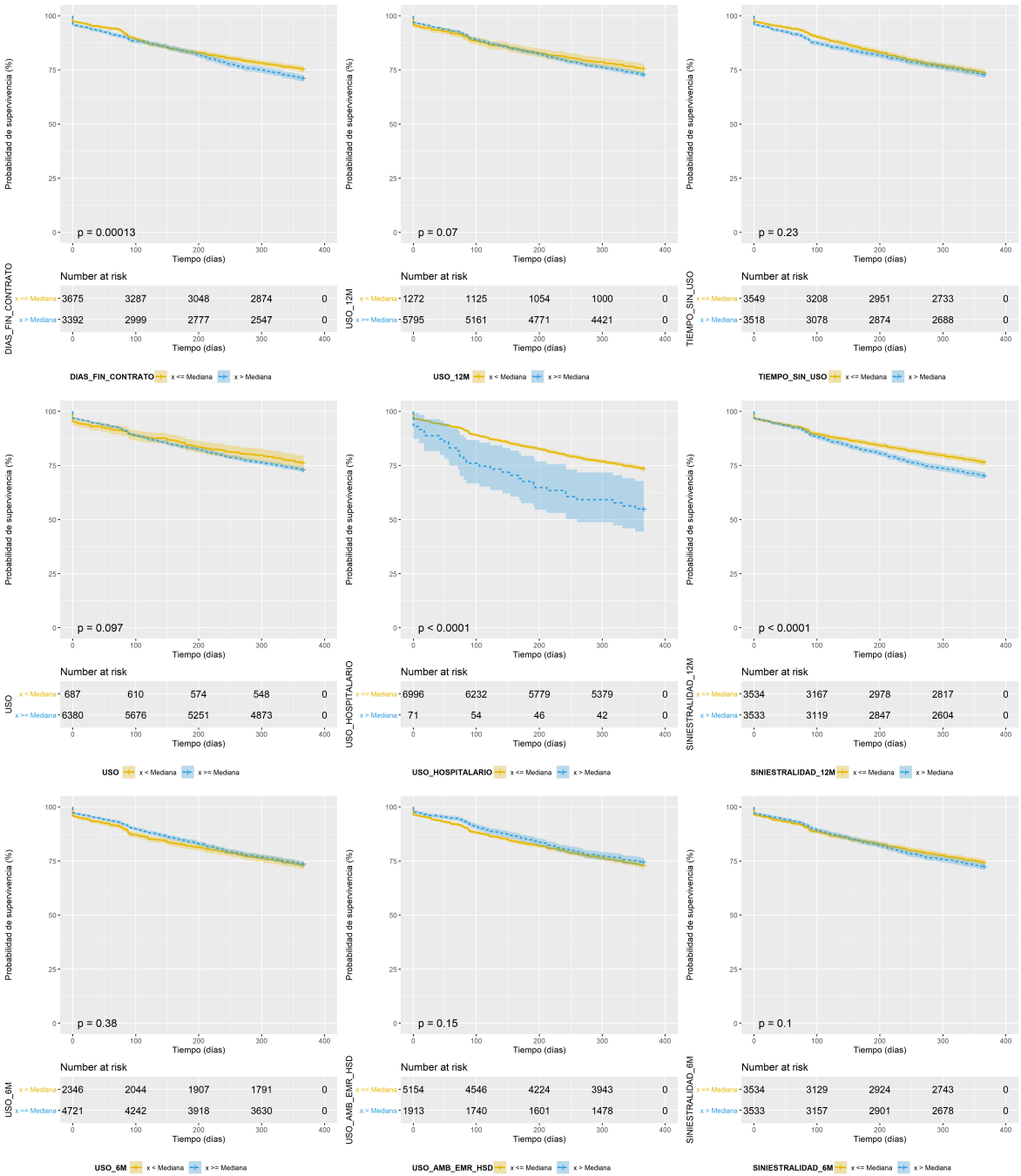


Figura 7.22: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 10

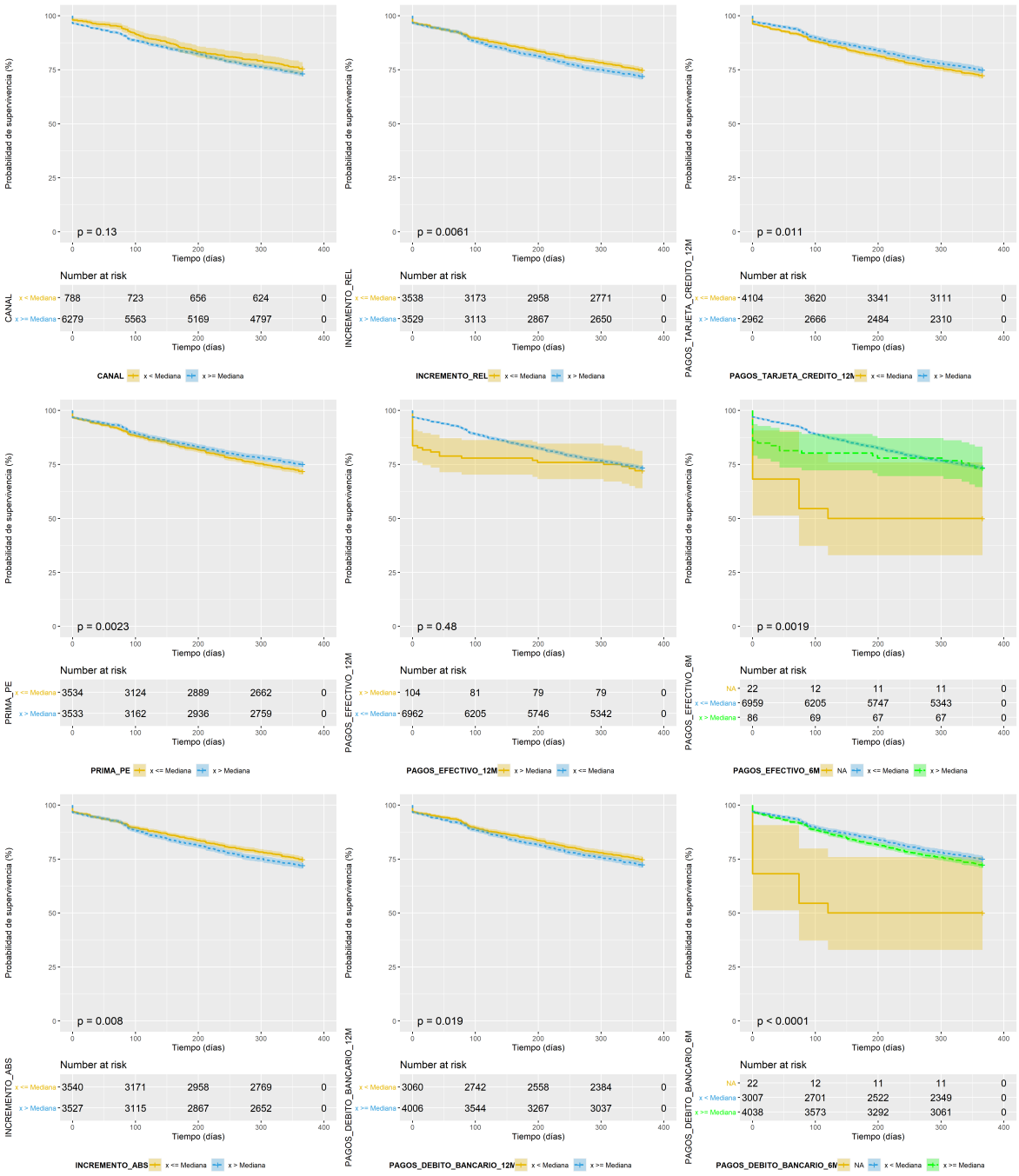


Figura 7.23: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 11

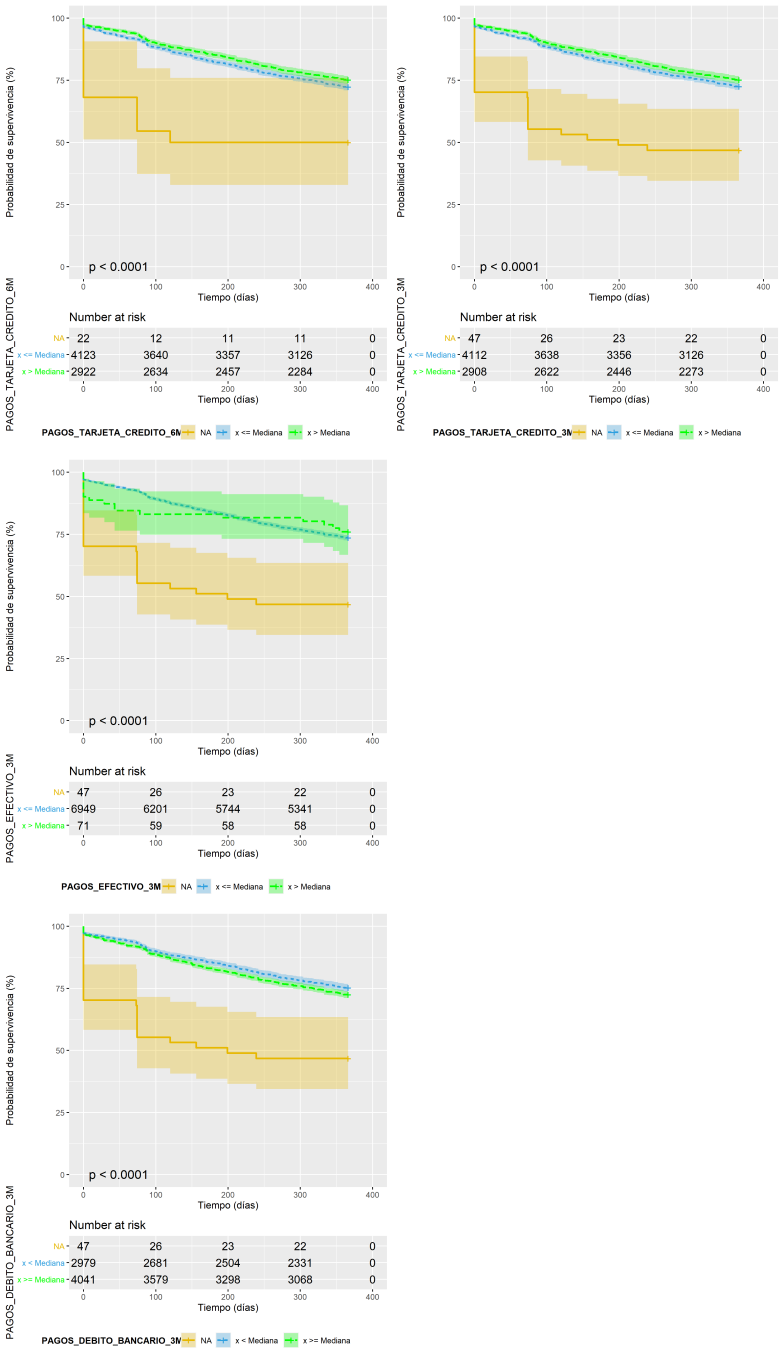


Figura 7.24: Población V2 - Estimador Kaplan y Meier y prueba Log-rank de la base de desarrollo partida por la mediana de cada variable, pag. 12

7.7 CÓDIGO R PARA MODELO DE COX POBLACIÓN V1

Código de R 7.2: Código R para modelo de población V1

```
################################################################################
################## Tesis Modelo de Regresión de Cox V1
##################     Autor: Marcos Pérez
################################################################################

############## Bibliotecas

library(readxl)
library(splines)
library(car)
library(sandwich)
library(effects)
library(data.table)
library(RcmdrMisc)
library(lubridate)
library(cluster)
library(dplyr)
library(fitdistrplus)
library(tcltk)
library(survival)
library(survminer)
library(ggplot2)

library(gridExtra)
library(grid)
library(lattice)
library(ggfortify)

library(riskRegression)
library(stargazer)


################################################################################

############## Base Cox y Base Testeo

load("Base_Cox.Rdata")
```

```
#load("C:/Users/marco.perez/Dropbox/Tema de Modelo Regresión de Cox/Modelo Cox/
    Bases Desarrollo y Prueba V1.Rdata")



################################################################################

############## Modelo de Cox 1

Modelo_Cox1 <- coxph(formula = Surv(TIEMPO, CENSURA) ~ CANAL + EDAD_TITULAR +
    vector_1 + vector_2 + vector_5 + vector_36 + vector_44 + vector_71,
    data = Base_Cox, y = TRUE, x = TRUE)

cox.zph(Modelo_Cox1)


################################################################################

############## Evaluación de Modelo de Cox 1

png(file = "Evaluación de Modelo de Cox V1.png", bg = "transparent",width = 15.5,
     height = 17, units = "cm", res= 200)
ggforest(Modelo_Cox1, data=Base_Cox, main="Evaluación de Modelo de Cox 1")
dev.off()



################################################################################

############## Gráfico de Modelo de Cox 1 Ajustado

cox_fit <- survfit(Modelo_Cox1)

png(file = "Modelo de Cox Ajustado V1.png", bg = "transparent",width = 15.5,
    height = 12, units = "cm", res= 200)
autoplot(cox_fit)+ labs( title="Modelo de Cox 1 Ajustado", x="Tiempo (días)",
    y="Probabilidad de Supervivencia") + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5))
dev.off()


```

```
-
- ##############################################################################
-
80 ############## Comportamiento respecto al tiempo de covariables de Modelo de Cox
       1
-
- aa_fit <-aareg( Surv(TIEMPO, CENSURA) ~ CANAL + EDAD_TITULAR +
-     vector_1 + vector_2 + vector_5 + vector_36 + vector_44 + vector_71,
-     data = Base_Cox)
85 aa_fit
-
- png(file = "Covariables respecto al tiempo V1.png", bg = "transparent",width =
       15.5, height = 12, units = "cm", res= 200)
- autoplot(aa_fit)+ labs( title="Modelo de Aalen para Covariables del Modelo de Cox
       1", x="Tiempo",
-     y="Riesgo Acumulativo") + theme(legend.key = element_blank(),
90     plot.title = element_text(hjust = 0.5))
- dev.off()
-
-
-
95 ##############################################################################
-
- ############## Gráficos de Verificación de Hipótesis de Riesgos Proporcionales
-
- png(file = "Verificación Hipótesis de Riesgo Proporcional V1.png", bg = "
       transparent",width = 15.5, height = 15.5, units = "cm", res= 200)
100 ggpar(ggcoxzph(cox.zph(Modelo_Cox1), point.size=.25, point.col="blue"),
-     font.main=c(6,"bold","black"),font.x=c(5,"plain","black"),
-     font.y=c(5,"plain","black"),font.tickslab=c(5,"plain","black"))
- dev.off()
-
105
-
- ##############################################################################
-
- ############## Verificación de Forma Funcional - Residuos de Martingala (
       Martigale))
110
- Base_Cox$resid_mart <- residuals(Modelo_Cox1, type = "martingale")
-
```

```
x1 <- ggpar( ggplot(data =  Base_Cox, mapping = aes(x = EDAD_TITULAR, y =
    resid_mart)) +
    geom_point(size=.25,col="blue") + geom_smooth(size=.25,col="#2E9FDF") + labs(
        title="Residuos de la Martingala vs EDAD_TITULAR",
    x="EDAD_TITULAR", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x2 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(CANAL), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs CANAL",
    x="CANAL", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x3 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_1), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_1",
    x="vector_1", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x4 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_2), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_2",
    x="vector_2", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

ggsave("Residuos de Martingala V1 pag.1.png",grid.arrange(x1,x2,x3,x4, ncol=2),
    width = 15.5, height = 15.5, units =  "cm")
```

```
-
-
- x1 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_5), y =
    resid_mart)) +
150     geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_5",
-     x="vector_5", y="Residuos de Martingala") +
-     theme_bw() + theme(legend.key = element_blank(),
-     plot.title = element_text(hjust = 0.5)),
-     font.main=c(8,"bold","black"),
155     font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))
-
- x2 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_36), y =
    resid_mart)) +
-     geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_36",
-     x="vector_36", y="Residuos de Martingala") +
160     theme_bw() + theme(legend.key = element_blank(),
-     plot.title = element_text(hjust = 0.5)),
-     font.main=c(8,"bold","black"),
-     font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))
-
165 x3 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_44), y =
    resid_mart)) +
-     geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_44",
-     x="vector_44", y="Residuos de Martingala") +
-     theme_bw() + theme(legend.key = element_blank(),
-     plot.title = element_text(hjust = 0.5)),
170     font.main=c(8,"bold","black"),
-     font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))
-
- x4 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_71), y =
    resid_mart)) +
-     geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_71",
175     x="vector_71", y="Residuos de Martingala") +
-     theme_bw() + theme(legend.key = element_blank(),
-     plot.title = element_text(hjust = 0.5)),
-     font.main=c(8,"bold","black"),
-     font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))
180
```

```
ggsave("Residuos de Martingala V1 pag.2.png",grid.arrange(x1,x2,x3,x4, ncol=2),
    width = 15.5, height = 15.5, units =  "cm")



ggcoxfunctional( Surv(TIEMPO, CENSURA) ~ CANAL + EDAD_TITULAR +
    vector_1 + vector_2 + vector_5 + vector_36 + vector_44 + vector_71,
    data = Base_Cox)


################################################################################

############## Verificación de que los sujetos no tienen influencia en la
############## estimación de cada coeficiente – Residuos de Puntajes (Score)

png(file = "Residuos de Puntajes V1.png", bg = "transparent", width = 15.5,
    height = 15.5, units = "cm", res= 400)
ggcoxdiagnostics(Modelo_Cox1, type = "score", point.size=.1, hline.size=.2,
    sline.size=.3, sline.lty=1) +
    labs( title="Residuos de Puntajes de Covariables del Modelo de Cox 1",
    x="Id de Observación", y="Residuos de Puntajes (score)") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
dev.off()

png(file = "Residuos de dfbeta V1.png", bg = "transparent", width = 15.5, height
    = 15.5, units = "cm", res= 400)
ggcoxdiagnostics(Modelo_Cox1, type = "dfbeta", point.size=.1, hline.size=.2,
    sline.size=.3, sline.lty=1) +
    labs( title="Residuos de dfbeta de Covariables del Modelo de Cox 1",
    x="Id de Observación", y="Residuos de dfbeta") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
dev.off()

png(file = "Residuos de dfbetas V1.png", bg = "transparent", width = 15.5, height
     = 15.5, units = "cm", res= 400)
ggcoxdiagnostics(Modelo_Cox1, type = "dfbetas", point.size=.1, hline.size=.2,
    sline.size=.3, sline.lty=1) +
    labs( title="Residuos de dfbetas de Covariables del Modelo de Cox 1",
    x="Id de Observación", y="Residuos de dfbetas") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
```

```
dev.off()



################################################################################

############## Que los sujetos no tienen influencia en la estimación del modelo
############## (No existen casos atípicos que hagan que el modelo sufra desvios)
############## – Residuos de Desvíos (Desviances)

png(file = "Residuos de Desvíos V1.png", bg = "transparent", width = 15.5, height
    = 15.5, units = "cm", res= 400)
ggcoxdiagnostics(Modelo_Cox1, type = "deviance", point.size=.15, hline.size=.25,
    sline.size=.5, sline.lty=1) +
    labs( title="Residuos de Desvíos del Modelo de Cox 1",
    x="Predicción Lineal", y="Residuos de Desvíos (deviance)") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
dev.off()



################################################################################

############## Base de Prueba V1

attach(Base_Prueba_V1)
Base_Prueba_V1$vector_1 <- ifelse(  PRESENTADO_6M_PA > 491.15   , 0, 1)
Base_Prueba_V1$vector_2 <- ifelse(  RECORTE_3M > 2  , 0, 1)
Base_Prueba_V1$vector_5 <- ifelse(  EDAD_TITULAR > 70  , 0, 1)
Base_Prueba_V1$vector_36 <- ifelse( CAMBIO_PLAN_12M > 0 &
    PAGOS_DEBITO_BANCARIO_12M > .857  , 1, 0)
Base_Prueba_V1$vector_44 <- ifelse( INDIVIDUAL_FAMILIAR > 1 &
    PAGOS_DEBITO_BANCARIO_12M > .857  , 1, 0)
Base_Prueba_V1$vector_71 <- ifelse( NUMERO_AFILIADOS_5ANOS_O_MENOS > 0 &
    TOPE_PLAN <= 30000  , 1, 0)
detach(Base_Prueba_V1)



################################################################################

############## Evaluación de Discriminación y Predicción por Base de Prueba V1
```

```
-
- library(riskRegression)
260 library(survC1)
-
- attach(Base_Prueba_V1)
- BP <- data.frame(TIEMPO,CENSURA,CANAL,EDAD_TITULAR,vector_1,vector_2,vector_5,
    vector_36,
-     vector_44,vector_71)
265 BPcsc <- data.frame(TIEMPO,CENSURA=1+CENSURA,CANAL,EDAD_TITULAR,vector_1,vector_2
    ,vector_5,vector_36,
-     vector_44,vector_71)
- detach(Base_Prueba_V1)
-
- EvalByRisk_V1 <- Est.Cval(mydata=cbind(time=BP$TIEMPO, event=BP$CENSURA,
270   predictions=predict(object=Modelo_Cox1, newdata=BP, type="risk")),
-   tau=370, nofit=TRUE)
-
- EvalByRisk_V1$Dhat
-
275
-
- ################################################################################
-
- ############## Validación Discriminación AUC y Brier Tiempo Dependiente Prueba V1
280
- score <- Score(list("MC1" = Modelo_Cox1),
-     formula = Surv(TIEMPO, CENSURA) ~ 1, data = BP,
-     times = seq(10,360,10), plots = "calibration", summary = "risks")
-
285 png(file = "AUC y Brier Prueba V1.png", bg = "transparent", width = 18, height =
    21, units = "cm", res= 400)
- par(mfrow=c(2,2))
- plotCalibration(score,time=90, xlab = "Riesgo Predicho", ylab = "Frecuncia
    Observada")
- text(x=.75, y = .1, "TIEMPO = 90 días", col = "blue", font=2)
- plotCalibration(score,time=180, xlab = "Riesgo Predicho", ylab = "Frecuncia
    Observada")
290 text(x=.75, y = .1, "TIEMPO = 180 días", col = "blue", font=2)
- plotCalibration(score,time=270, xlab = "Riesgo Predicho", ylab = "Frecuncia
    Observada")
- text(x=.75, y = .1, "TIEMPO = 270 días", col = "blue", font=2)
```

```
plotCalibration(score,time=360, xlab = "Riesgo Predicho", ylab = "Frecuncia
    Observada")
text(x=.75, y = .1, "TIEMPO = 180 días", col = "blue", font=2)
dev.off()

ggplot(data = score$AUC$score, aes(x = times, y = AUC)) + ###, colour = model)) +
    geom_point(col = "blue") + geom_line(col = "#2E9FDF") + geom_ribbon(data =
        score$AUC$score,
    aes(ymin = lower, ymax = upper), alpha=.1) + labs(x = "Tiempo (días)",
    title = "AUC Tiempo-Dependiente en Prueba Modelo de Cox 1") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
ggsave("AUC tiempo-dependiente Prueba V1.png", width = 15.5, height = 15.5, units
     =  "cm")

ggplot(data = score$Brier$score, aes(x = times, y = Brier, colour = model)) +
    geom_point() + geom_line() + geom_ribbon(data = score$Brier$score,
    aes(ymin = lower, ymax = upper, colour = model), alpha=.1, linetype = 2) +
    labs(x = "Tiempo (días)", title = "Brier Tiempo-Dependiente en Prueba Modelo
        de Cox 1") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
ggsave("Brier tiempo-dependiente Prueba V1.png", width = 15.5, height = 15.5,
    units =  "cm")



################################################################################

############## Nomogramas

library(regplot)

png(file = "Nomograma V1 - Riesgo Alto.png", bg = "transparent", width = 15.5,
    height = 15.5, units = "cm", res= 400)
regplot(Modelo_Cox1,observation=BP[1,], clickable=TRUE,
points=TRUE, rank="sd",failtime = c(180,360))
dev.off()

png(file = "Nomograma V1 - Riesgo Medio.png", bg = "transparent", width = 15.5,
    height = 15.5, units = "cm", res= 400)
regplot(Modelo_Cox1,observation=BP[17,], clickable=TRUE,
points=TRUE, rank="sd",failtime = c(180,360))
dev.off()
```

```
-
- png(file = "Nomograma V1 – Riesgo Bajo.png", bg = "transparent", width = 15.5,
      height = 15.5, units = "cm", res= 400)
330 regplot(Modelo_Cox1,observation=BP[6,], clickable=TRUE,
- points=TRUE, rank="sd",failtime = c(180,360))
- dev.off()
```

7.8 CÓDIGO R PARA MODELO DE COX POBLACIÓN V2

Código de R 7.3: Código R para modelo de población V2

```
################################################################################
################## Tesis Modelo de Regresión de Cox V2
##################     Autor: Marcos Pérez
################################################################################

############## Bibliotecas

library(readxl)
library(splines)
library(car)
library(sandwich)
library(effects)
library(data.table)
library(RcmdrMisc)
library(lubridate)
library(cluster)
library(dplyr)
library(fitdistrplus)
library(tcltk)
library(survival)
library(survminer)
library(ggplot2)

library(gridExtra)
library(grid)
library(lattice)
library(ggfortify)

library(riskRegression)
library(stargazer)

################################################################################

############## Base Cox y Base Testeo

load("Base_Cox.Rdata")
```

```
#load("C:/Users/marco.perez/Dropbox/Tema de Modelo Regresión de Cox/Modelo Cox/
    Bases Desarrollo y Prueba V1.Rdata")
-
-
40
-
-
################################################################################
-
-############## Modelo de Cox 2
45
-Modelo_Cox2 <- coxph(formula = Surv(TIEMPO, CENSURA) ~ MORA_MAX_6M +
-    vector_2  + vector_27 + vector_30 + vector_37 + vector_50 + vector_52 +
-    vector_55 + vector_61 + vector_68 + vector_72 + vector_73 + vector_75,
-    data = Base_Cox, y = TRUE, x = TRUE)
50
-cox.zph(Modelo_Cox2)
-
-
-
55################################################################################
-
-############## Evaluación de Modelo de Cox 2
-
-png(file = "Evaluación de Modelo de Cox V2.png", bg = "transparent",width = 15.5,
     height = 17, units = "cm", res= 200)
60ggforest(Modelo_Cox2, data=Base_Cox, main="Evaluación de Modelo de Cox 2")
-dev.off()
-
-
-
65################################################################################
-
-############## Gráfico de Modelo de Cox 2 Ajustado
-
-cox_fit <- survfit(Modelo_Cox2)
70
-png(file = "Modelo de Cox Ajustado V2.png", bg = "transparent",width = 15.5,
    height = 12, units = "cm", res= 200)
-autoplot(cox_fit)+ labs( title="Modelo de Cox 2 Ajustado", x="Tiempo (días)",
-    y="Probabilidad de Supervivencia") + theme(legend.key = element_blank(),
-    plot.title = element_text(hjust = 0.5))
75dev.off()
```

```
-
-
-
-  ##############################################################################
80
-  ############## Comportamiento respecto al tiempo de covariables de Modelo de Cox
       2
-
-  aa_fit <-aareg( Surv(TIEMPO, CENSURA) ~ MORA_MAX_6M +
-      vector_2  + vector_27 + vector_30 + vector_37 + vector_50 + vector_52 +
85    vector_55 + vector_61 + vector_68 + vector_72 + vector_73 + vector_75,
-     data = Base_Cox)
-  aa_fit
-
-  png(file = "Covariables respecto al tiempo V2.png", bg = "transparent",width =
       20.5, height = 16.5, units = "cm", res= 200)
90  autoplot(aa_fit)+ labs( title="Modelo de Aalen para Covariables del Modelo de Cox
       2", x="Tiempo",
-     y="Riesgo Acumulativo") + theme(legend.key = element_blank(),
-     plot.title = element_text(hjust = 0.5))
-  dev.off()
-
95
-
-  ##############################################################################
-
-  ############## Gráficos de Verificación de Hipótesis de Riesgos Proporcionales
100
-  png(file = "Verificación Hipótesis de Riesgo Proporcional V2.png", bg = "
       transparent",width = 15.5, height = 15.5, units = "cm", res= 200)
-  ggpar(ggcoxzph(cox.zph(Modelo_Cox2), point.size=.25, point.col="blue"),
-      font.main=c(5,"bold","black"),font.x=c(5,"plain","black"),
-      font.y=c(5,"plain","black"),font.tickslab=c(4,"plain","black"))
105 dev.off()
-
-
-
-  ##############################################################################
110
-  ############## Verificación de Forma Funcional – Residuos de Martingala (
       Martigale))
-
```

```
Base_Cox$resid_mart <- residuals(Modelo_Cox2, type = "martingale")

x1 <- ggpar( ggplot(data =  Base_Cox, mapping = aes(x = MORA_MAX_6M, y =
    resid_mart)) +
    geom_point(size=.25,col="blue") + geom_smooth(size=.25,col="#2E9FDF") + labs(
         title="Residuos de la Martingala vs MORA_MAX_6M",
    x="MORA_MAX_6M", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x2 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_2), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_2",
    x="vector_2", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x3 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_27), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_27",
    x="vector_27", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x4 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_30), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_30",
    x="vector_30", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))
```

```
ggsave("Residuos de Martingala V2 pag.1.png",grid.arrange(x1,x2,x3,x4, ncol=2),
    width = 15.5, height = 15.5, units =  "cm")


x1 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_37), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_37",
    x="vector_37", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x2 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_50), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_50",
    x="vector_50", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x3 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_52), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_52",
    x="vector_52", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x4 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_55), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_55",
    x="vector_55", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
```

```
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

ggsave("Residuos de Martingala V2 pag.2.png",grid.arrange(x1,x2,x3,x4, ncol=2),
    width = 15.5, height = 15.5, units =  "cm")



x1 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_61), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_61",
    x="vector_61", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x2 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_68), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_68",
    x="vector_68", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x3 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_72), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_72",
    x="vector_72", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

x4 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_73), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_73",
    x="vector_73", y="Residuos de Martingala") +
```

```
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

ggsave("Residuos de Martingala V2 pag.3.png",grid.arrange(x1,x2,x3,x4, ncol=2),
    width = 15.5, height = 15.5, units =  "cm")



x1 <- ggpar( ggplot(data = Base_Cox, mapping = aes(x = ordered(vector_75), y =
    resid_mart)) +
    geom_violin(size=.4,col="#2E9FDF") + geom_point(size=.5,col="blue") + labs(
        title="Residuos de Martingala vs vector_75",
    x="vector_75", y="Residuos de Martingala") +
    theme_bw() + theme(legend.key = element_blank(),
    plot.title = element_text(hjust = 0.5)),
    font.main=c(8,"bold","black"),
    font.x=c(8,"plain","black"),font.y=c(8,"plain","black"))

ggsave("Residuos de Martingala V2 pag.4.png",grid.arrange(x1, ncol=2),
    width = 15.5, height = 15.5, units =  "cm")



#ggcoxfunctional( Surv(TIEMPO, CENSURA) ~ CANAL + EDAD_TITULAR +
#    vector_1 + vector_2 + vector_5 + vector_36 + vector_44 + vector_71,
#    data = Base_Cox)



################################################################################

############## Verificación de que los sujetos no tienen influencia en la
############## estimación de cada coeficiente – Residuos de Puntajes (Score)

png(file = "Residuos de Puntajes V2.png", bg = "transparent", width = 15.5,
    height = 15.5, units = "cm", res= 400)
ggcoxdiagnostics(Modelo_Cox2, type = "score", point.size=.1, hline.size=.2,
    sline.size=.3, sline.lty=1) +
    labs( title="Residuos de Puntajes de Covariables del Modelo de Cox 2",
    x="Id de Observación", y="Residuos de Puntajes (score)") +
```

```
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
dev.off()

png(file = "Residuos de dfbeta V2.png", bg = "transparent", width = 15.5, height
    = 15.5, units = "cm", res= 400)
ggcoxdiagnostics(Modelo_Cox2, type = "dfbeta", point.size=.1, hline.size=.2,
    sline.size=.3, sline.lty=1) +
    labs( title="Residuos de dfbeta de Covariables del Modelo de Cox 2",
    x="Id de Observación", y="Residuos de dfbeta") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
dev.off()

png(file = "Residuos de dfbetas V2.png", bg = "transparent", width = 15.5, height
     = 15.5, units = "cm", res= 400)
ggcoxdiagnostics(Modelo_Cox2, type = "dfbetas", point.size=.1, hline.size=.2,
    sline.size=.3, sline.lty=1) +
    labs( title="Residuos de dfbetas de Covariables del Modelo de Cox 2",
    x="Id de Observación", y="Residuos de dfbetas") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
dev.off()



################################################################################

############## Que los sujetos no tienen influencia en la estimación del modelo
############## (No existen casos atípicos que hagan que el modelo sufra desvios)
############## - Residuos de Desvíos (Desviances)

png(file = "Residuos de Desvíos V2.png", bg = "transparent", width = 15.5, height
     = 15.5, units = "cm", res= 400)
ggcoxdiagnostics(Modelo_Cox2, type = "deviance", point.size=.15, hline.size=.25,
    sline.size=.5, sline.lty=1) +
    labs( title="Residuos de Desvíos del Modelo de Cox 2",
    x="Predicción Lineal", y="Residuos de Desvíos (deviance)") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
dev.off()



################################################################################

```

```
############## Base de Prueba V2

attach(Base_Prueba_V2)
Base_Prueba_V2$vector_2 <- ifelse(  RECORTE_3M <= 61.61  , 0, 1)
Base_Prueba_V2$vector_27 <- ifelse( MORA_MAX_6M <= 6 & INCREMENTO_REL <= .156  ,
    0, 1)
Base_Prueba_V2$vector_30 <- ifelse( EDAD_TITULAR <= 49 & TIEMPO_AFILIACION <=
    1484  , 1, 0)
Base_Prueba_V2$vector_37 <- ifelse( RECORTE_6M <= 120 & MORA_MAX_12M <= 31  , 0,
    1)
Base_Prueba_V2$vector_50 <- ifelse( EDAD_TITULAR <= 39 & INCREMENTO_ABS > 12.97  ,
     1, 0)
Base_Prueba_V2$vector_52 <- ifelse( PRODUCTO_MH > 0 & MORA_MAX_3M <= 14  , 0, 1)
Base_Prueba_V2$vector_55 <- ifelse( PRESENTADO_CPC_12M <= 328.62 &
    SINIESTRALIDAD_12M <= .513 & SINIESTRALIDAD_12M > .116  , 1, 0)
Base_Prueba_V2$vector_61 <- ifelse( SINIESTRALIDAD_12M > .513 & INCREMENTO_ABS >
    6.72  , 1, 0)
Base_Prueba_V2$vector_68 <- ifelse( NUMERO_AFILIADOS_2ANOS_O_MENOS <= 0 &
    SINIESTRALIDAD_6M <= .763  , 0, 1)
Base_Prueba_V2$vector_72 <- ifelse( NUMERO_AFILIADOS_MENORES > 1 & INCREMENTO_REL
     > .156  , 1, 0)
Base_Prueba_V2$vector_73 <- ifelse( PRESENTADO_CRONICO_12M > 619.28 &
    SINIESTRALIDAD_12M <= .513  , 0, 1)
Base_Prueba_V2$vector_75 <- ifelse( EDAD_TITULAR <= 49 & DIAS_FIN_CONTRATO > 318
    , 1, 0)
detach(Base_Prueba_V2)



################################################################################

############## Evaluación de Discriminación y Predicción por Base de Prueba V2

library(riskRegression)
library(survC1)

attach(Base_Prueba_V2)
BP <- na.omit(data.frame(TIEMPO,CENSURA,MORA_MAX_6M, vector_2 , vector_27 , vector_30 ,
    vector_37 , vector_50 , vector_52 , vector_55 , vector_61 , vector_68 , vector_72 ,
    vector_73 , vector_75))
BPcsc <- na.omit(data.frame(TIEMPO,CENSURA=1+CENSURA,MORA_MAX_6M, vector_2 ,
    vector_27 ,
```

```
    vector_30,vector_37,vector_50,vector_52,vector_55,vector_61,vector_68,
    vector_72,vector_73,vector_75))
detach(Base_Prueba_V2)

EvalByRisk_V2 <- Est.Cval(mydata=cbind(time=BP$TIEMPO, event=BP$CENSURA,
  predictions=predict(object=Modelo_Cox2, newdata=BP, type="risk")),
  tau=370, nofit=TRUE)

EvalByRisk_V2$Dhat


################################################################################

############## Validación Discriminación AUC y Brier Tiempo Dependiente Prueba V2

score <- Score(list("MC2" = Modelo_Cox2),
    formula = Surv(TIEMPO, CENSURA) ~ 1, data = BP,
    times = seq(10,360,10), plots = "calibration", summary = "risks")

png(file = "AUC y Brier Prueba V2.png", bg = "transparent", width = 18, height =
    21, units = "cm", res= 400)
par(mfrow=c(2,2))
plotCalibration(score,time=90, xlab = "Riesgo Predicho", ylab = "Frecuncia
    Observada")
text(x=.75, y = .1, "TIEMPO = 90 días", col = "blue", font=2)
plotCalibration(score,time=180, xlab = "Riesgo Predicho", ylab = "Frecuncia
    Observada")
text(x=.75, y = .1, "TIEMPO = 180 días", col = "blue", font=2)
plotCalibration(score,time=270, xlab = "Riesgo Predicho", ylab = "Frecuncia
    Observada")
text(x=.75, y = .1, "TIEMPO = 270 días", col = "blue", font=2)
plotCalibration(score,time=360, xlab = "Riesgo Predicho", ylab = "Frecuncia
    Observada")
text(x=.75, y = .1, "TIEMPO = 180 días", col = "blue", font=2)
dev.off()

ggplot(data = score$AUC$score, aes(x = times, y = AUC)) + ###, colour = model)) +
    geom_point(col = "blue") + geom_line(col = "#2E9FDF") + geom_ribbon(data =
        score$AUC$score,
    aes(ymin = lower, ymax = upper), alpha=.1) + labs(x = "Tiempo (días)",
    title = "AUC Tiempo-Dependiente en Prueba Modelo de Cox 2") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
```

```
ggsave("AUC tiempo-dependiente Prueba V2.png", width = 15.5, height = 15.5, units
    = "cm")

ggplot(data = score$Brier$score, aes(x = times, y = Brier, colour = model)) +
    geom_point() + geom_line() + geom_ribbon(data = score$Brier$score,
    aes(ymin = lower, ymax = upper, colour = model), alpha=.1, linetype = 2) +
    labs(x = "Tiempo (días)", title = "Brier Tiempo-Dependiente en Prueba Modelo
        de Cox 2") +
    theme(legend.key = element_blank(), plot.title = element_text(hjust = 0.5))
ggsave("Brier tiempo-dependiente Prueba V2.png", width = 15.5, height = 15.5,
    units = "cm")



################################################################################

############## Nomogramas

library(regplot)

png(file = "Nomograma V2 - Riesgo Alto.png", bg = "transparent", width = 15.5,
    height = 15.5, units = "cm", res= 400)
regplot(Modelo_Cox2,observation=BP[2305,], clickable=TRUE,
points=TRUE, rank="sd",failtime = c(180,360))
dev.off()

png(file = "Nomograma V2 - Riesgo Medio.png", bg = "transparent", width = 15.5,
    height = 15.5, units = "cm", res= 400)
regplot(Modelo_Cox2,observation=BP[2235,], clickable=TRUE,
points=TRUE, rank="sd",failtime = c(180,360))
dev.off()

png(file = "Nomograma V2 - Riesgo Bajo.png", bg = "transparent", width = 15.5,
    height = 15.5, units = "cm", res= 400)
regplot(Modelo_Cox2,observation=BP[8,], clickable=TRUE,
points=TRUE, rank="sd",failtime = c(180,360))
dev.off()
```