UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

ESTIMANDO UNA ESTRUCTURA DE PROBABILIDADES DE INCUMPLIMIENTO CREDITICIO PARA UNA CARTERA DE CONSUMO, MEDIANTE ANÁLISIS DE SUPERVIVENCIA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JORGE IGNACIO AVENDAÑO MATURANA

PROFESOR GUÍA:
JOSÉ MIGUEL CRUZ GONZÁLEZ

MIEMBROS DE LA COMISIÓN:
LUIS ABURTO LAFOURCADE
RICHARD WEBER HAAS

Este trabajo ha sido parcialmente financiado por Banco Bci

SANTIAGO DE CHILE
2017

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: JORGE IGNACIO AVENDAÑO MATURANA
FECHA: 2017
PROF. GUÍA: SR. JOSÉ MIGUEL CRUZ GONZÁLEZ

ESTIMANDO UNA ESTRUCTURA DE PROBABILIDADES DE INCUMPLIMIENTO CREDITICIO PARA UNA CARTERA DE CONSUMO, MEDIANTE ANÁLISIS DE SUPERVIVENCIA

Una entidad bancaria nacional propuso el problema de pronosticar probabilidades de incumplimiento crediticio a lo largo del tiempo (PI), motivada por cambios en estándares internacionales y por buscar una visión extendida del riesgo de sus clientes de consumo.

Dentro de los objetivos del proyecto destacan: determinar una expresión para la PI como función del tiempo, evaluar y discriminar los mejores modelos de acuerdo a criterios estadísticos y operativos, y crear curvas específicas para diferentes segmentos de clientes.

El enfoque utilizado para la predicción fue el análisis de supervivencia, el cual modela de forma probabilística el tiempo hasta que ocurre un incumplimiento. Para ello se contó con un panel de 10.000 clientes, el cual consigna para cada uno de ellos, su tiempo de supervivencia, su estatus de incumplimiento e información longitudinal de variables tales como: la mora, la antigüedad de la cuenta corriente, el score crediticio y el estatus de renegociación.

La metodología consistió en tres etapas: pre-procesamiento y transformación, exploración y procesamiento de datos. En la primera se aplicaron los filtros de datos, en la segunda se buscaron patrones en ellos y en la tercera etapa se entrenaron los modelos.

Se evaluaron los modelos mediante criterios estadísticos y operativos. En los primeros figuran el criterio de información de Aikake y los residuos de Cox-Snell. Mientras que los segundos se basaron en la adaptación de los modelos a los procesos originación (ingreso de nuevos clientes) y seguimiento de clientes (evaluación a lo largo del tiempo).

Los resultados indican que la mora y la renegociación actúan en favor del incumplimiento, mientras que el score y la antigüedad lo hacen en contra. Esto se cumplió para la mayoría de los modelos con un 99 % de confianza.

Dentro de las conclusiones se destaca que el mejor modelo de originación es la regresión AFT lognormal con covariables evaluadas al inicio, porque la distribución de sus residuos fue la más cercana a una exponencial de tasa 1 y porque evalúa a nuevos clientes de acuerdo a un modelo entrenado con individuos de su misma condición. Mientras que el mejor modelo de seguimiento corresponde a una regresión de Cox con covariables dependientes del tiempo, ya que incorpora la historia de los clientes y fue el modelo con menor AIC dentro de los modelos semi-parametricos.

Finalmente se destaca que el incumplimiento crediticio se puede interpretar como un fenómeno de supervivencia, cuyo enfoque también admite extensiones, tales como la incorporación del pago temprano como nuevo evento y la inclusión de variables macroeconómicas. Estos aspectos se propondrán para futuros desarrollos.

A Lorena, Zunita y Pablo. Este logro es por ustedes.

Tabla de contenido

Índice de tablas

Índice de ilustraciones

Capítulo 1

Introducción

La probabilidad de incumpliemiento (PI) es fundamental para la determinación del riesgo de crédito, el cual se define como el potencial que un cliente o contraparte incumpla sus obligaciones según los términos acordados (Basel Committee On Banking Supervision, 2000). La correcta evaluación del riesgo es relevante, ya que su mal manejo puede deteriorar la situación financiera de la entidad y, por consiguiente, derivar en el incumplimiento de sus propias obligaciones.

Las aplicaciones de la PI son variadas, entre las cuales destacan la asignación de crédito a nuevos clientes, el seguimiento a clientes ya existentes, la fijación de las provisiones bancarias y la fijación de las tasas de interés a cobrar por crédito, entre otras.

La práctica más extendida para calcular las probabilidades de incumplimiento sobre carteras de créditos de consumo, corresponde al método de *credit scoring*.

Este método típicamente se basa en el ajuste de un modelo de regresión logística, el cual incorpora características de los clientes y busca predecir una probabilidad teórica de incumplir en los próximos 12 meses. Sin embargo, para efectos de gestión, esta probabilidad teórica (con valores entre 0 y 1) se reescala para obtener un *score*, el cual posee un mayor rango de variación y se interpreta de modo que un mayor valor refleja que el cliente es de mejor calidad. Así, la clasificación entre buenos y malos pagadores se alcanza al comparar este valor con un nivel de corte definido por la entidad bancaria (Andreeva, 2006).

Según la decisión de negocios de negocios a tomar, se distinguen dos tipos de *scoring*: el *application scoring*, el cual se utiliza para apoyar la entrega de créditos a nuevos clientes, y se sustenta en el modelo logístico antes descrito; y el *behavioural scoring*, empleado para evaluar el riesgo de clientes ya existentes (Thomas, 2000).

Dentro de la empresa, este tipo de modelos toman los nombres de modelos de originación y modelos de seguimiento, respectivamente. Siendo estas denominaciones las que se utlizarán en este trabajo de memoria.

En dicho contexto, el interés de la entidad está enfocado en buscar un nuevo modelo de seguimiento, pero con la particularidad de describir una curva de PI a lo largo de múltiples

periodos del tiempo.

La motivación se debe a la necesidad de anticiparse a los cambios que traerá consigo la aplicación del Estándar Internacional de Reportes Financieros (IFRS9), y en la actualización de las prácticas actuales de seguimiento.

En relación a IFRS9 (a aplicarse en 2018 en Europa), el estándar establecerá un nuevo marco para el cálculo de las pérdidas esperadas de crédito, el cual solicitará a los bancos reconocer el valor de vida de las pérdidas esperadas de crédito (*lifetime expected credit losses* o LECL) cuando los préstamos comiencen a deteriorarse (IFRS, 2014).

Dicho cambio es relevante para efectos de este trabajo, puesto que el LECL requerirá una PI de largo plazo que recree el riesgo de incumplimiento a lo largo de la vida del crédito, superando el intervalo de 12 meses que exige la regulación (Basel Committee On Banking Supervision, 2005)[1].

Por otro lado, para el seguimiento de clientes, actualmente la entidad ajusta un modelo de scoring basado en regresión logística. Con los resultados se obtiene el score de cada cliente y luego éstos son segmentados en grupos de acuerdo a dicho valor. Posteriormente, para cada segmento de riesgo, se calcula la razón entre el número de incumplimientos y el total de clientes a lo largo de la ventana de desempeño (1 a 12 meses más), para luego utilizar dicho valor como indicador de la PI.

Las desventajas del método actual se centran en el último paso y se resumen en los siguientes puntos:

- No se pueden establecer proyecciones de la PI que sean más largas que 12 meses de duración.
- Sólo se observa una probabilidad puntual y no una trayectoria que refleje la dinámica del incumplimiento a lo largo de la maduración del crédito.
- No se pueden desarrollar enfoques de cálculo de pérdida esperada en que la PI esté sincronizada con las diferentes cuotas de crédito.

De este modo, un método de seguimiento que permita describir las probabilidades de incumplimiento en el tiempo, permitiría lidiar con dichos problemas.

De acuerdo con Thomas (2000), los modelos de seguimiento se dividen en dos enfoques: aquellos que emplean el método del *application scoring* (es decir, modelos de regresión logística) y aquellos que construyen modelos probabilísticos para el comportamiento del cliente.

Dentro de estos últimos se han perfilado los modelos basados en análisis de supervivencia (AS), los cuales buscan modelar el tiempo hasta que ocurre el incumplimiento. De de las ventajas expuestas por Banasik, Crook, y Thomas (1999) en torno a los modelos de supervivencia, se destacan las siguientes:

- La posibilidad de recerar el proceso de incumplimiento, permitiendo la presencia de

[1]Para mayores detalles revise el Capítulo 10.

casos en los cuales no se ha observado el incumplimiento al final del periodo[2].

- Proveer un pronóstico como función del tiempo en una simple ecuación.

Para modelos de originación, estudios también han empleado este enfoque (Banasik y cols., 1999; Bellotti y Crook, 2009), pero bajo el objetivo de compararlo con el modelo de regresión logística tradicional. Por lo que sólo pronostican la probabilidad a 12 meses, a pesar de que el análisis de supervivencia brinda la posibilidad de pronosticar probabilidades para tiempos mayores. De todos modos, estos estudios han brindado resultados competitivos con respecto al modelo logístico.

Dada la búsqueda de estimar la trayectoria de la probabilidad de incumplimiento, esta memoria toma distancia respecto de estos dos últimos trabajos anteriores. No obstante, las herramientas matemáticas empleadas, tienen el potencial para la construcción de esta curva.

[2]Más adelante se verá que este fenómeno se denomina como *censura derecha*.

Capítulo 2

Objetivos y alcances

2.1. Objetivo general

El objetivo general corresponde a calibrar y pronosticar probabilidades de incumplimiento crediticio en clientes (PI), considerando múltiples periodos del tiempo y permitiendo tanto una visión refinada (de mes a mes) como de largo plazo.

2.2. Objetivos específicos

Este trabajo de memoria tiene como objetivos específicos: determinar una forma funcional para la PI en función del tiempo, evaluar y discriminar los mejores modelos de PI de acuerdo a criterios operativos y estadísticos, y crear curvas de PI específicas para diferentes segmentos de clientes.

2.3. Alcances

Los alcances de este trabajo de memoria están definidos por cinco condiciones, las cuales se describirán a continuación.

En primer lugar, no se buscará cambiar el modelo de scoring que ostenta actualmente la entidad bajo estudio, sino que sólo se cambiará la forma de calcular la PI post-scoring. En segundo lugar, no se contemplarán desarrollos sobre el proceso de pricing y el cálculo de las pérdidas esperadas, por lo que se privilegiará el estudio exclusivo de la PI. En tercer lugar, el estudio se focalizará en modelar las probabilidades asociadas a eventos de incumplimiento, dejando de lado otra clase de eventos, tales como: el pago temprano o el saneamiento de clientes que incumplieron anteriormente. En cuarto lugar, no se considerarán en los modelos los efectos de variables macroeconómicas.

Capítulo 3

Marco teórico

3.1. Enfoque de desarrollo

En este trabajo de memoria se optó por el enfoque de análisis de supervivencia para la construcción de las curvas de PI, puesto que permite realizar proyecciones para múltiples periodos, tanto para clientes individuales como para grupos de clientes. Además, este tipo de modelos es adecuado a la estructura presente en los datos de crédito de consumo.

Para ilustrar este punto punto, considere el siguiente ejemplo: suponga la presencia de un nuevo cliente, quien abrió un crédito de consumo en un tiempo calendario t. A lo largo de su vida se registra su información, para finalmente verificar el incumplimiento o el pago sus obligaciones en un tiempo $t+T$, donde T corresponde a la maduración o tiempo de vida del cliente.

La recopilación anterior se puede resumir de la siguiente forma:

$$\left\{T_{\mathrm{i}}, \delta_{\mathrm{i}}, \{Z_{\mathrm{i}}(t_{\mathrm{i}j})\}_{t_{\mathrm{i}j}=1}^{T_{\mathrm{i}}}\right\}_{\mathrm{i}=1}^{N} \tag{3.1}$$

Donde T_{i} corresponde al tiempo de supervivencia del cliente i (en un espacio de duración, no de tiempo calendario), δ_{i} es el estatus de incumplimiento al final de su vida (i.e. 1 si incumple y 0 si no) y $\{Z_{\mathrm{i}}(t_{\mathrm{i}j})\}_{t_{\mathrm{i}j}=1}^{T_{\mathrm{i}}}$ corresponde a la información adicional del cliente, extraída a lo largo de múltiples mediciones $t_{\mathrm{i}j}$ hasta T_{i}.

Este tipo de datos es característico en los fenómenos de supervivencia. Por lo tanto, el fenómeno del incumplimiento crediticio puede ser interpretado bajo esta óptica.

3.2. Análisis de supervivencia

El análisis de supervivencia corresponde a una metodología que busca modelar probabilísticamente el tiempo hasta que ocurre un evento (tiempo de falla). Dicho evento puede

ser la muerte, la ocurrencia de una enfermedad, como también el incumplimiento crediticio (que es el caso de estudio de esta memoria).

En efecto, se asume que se tiene una muestra de n individuos independientes, cuyo comportamiento es examinado en un intervalo de tiempo determinado. A medida que transcurren los periodos se van observando los eventos de incumplimiento y se registran.

Dentro del análisis de supervivencia, un problema principal examinado es el desarrollo de métodos para evaluar la dependencia del tiempo de falla con variables explicativas. Un segundo problema involucra la estimación y la especificación de modelos para la distrubución subyacente del tiempo de falla (Kalbfleisch y Prentice, 2002).

3.2.1. Aspectos elementales

Para construir un modelo de supervivencia, hay que asumir la existencia de una variable aleatoria T que describe el tiempo de falla y cumple la siguiente ley de probabilidad:

$$F(t) = \Pr(T \leq t) \tag{3.2}$$

Se define la función de supervivencia como:

$$S(t) = \Pr(T > t) = 1 - F(t) \tag{3.3}$$

Es decir, la probabilidad de que un cliente incumpla después de un periodo t.

Una forma alternativa de determinar la distribución de T es a partir de la función de riesgo (*hazard function*), definida como:

$$h(t) = \lim_{\mathrm{d}t \to 0} \Pr(t \leq T \leq t + \mathrm{d}t | T \geq t)/\mathrm{d}t \tag{3.4}$$

Una definición equivalente corresponde a:

$$h(t) = \frac{f(t)}{S(t)} \tag{3.5}$$

Donde $f(t)$ es la función densidad de probabilidad de T.

Se puede demostrar que:

$$h(t) = -\frac{\mathrm{d}}{\mathrm{d}t} \ln S(t) \tag{3.6}$$

Resolviendo la ecuación diferencial con condición inicial $S(0) = 1$, se obtiene que la función de supervivencia está dada por:

$$S(t) = \exp\left(-\int_0^t h(x)\mathrm{d}x\right) \tag{3.7}$$

Cabe señalar que estos modelos se estiman a través del método de máxima verosimilitud.

3.2.2. Censura

La censura es una característica peculiar en los estudios de supervivencia (y no ajeno al fenómeno del incumplimiento crediticio), la cual ocurre cuando se determina que algunos tiempos de vida ocurren sólo entre ciertos intervalos (Klein y Moeschberger, 2005). Por el contrario, los datos no censurados, son aquellos para los cuales se sabe exactamente dicho tiempo de vida.

Se reconocen diferentes tipos de censura: la censura derecha, en la cual se asume que el cliente incumplirá en un tiempo (T_{i}, ∞), pero no cuándo exactamente; la censura izquierda, asociada a clientes que ya experimentaron el incumplimiento antes de que hayan sido observados por el estudio; y la censura por intervalos, que se asume en estudios con clientes que presentan tanto censura derecha como izquierda (además de aquellos con sus tiempos exactamente medidos).

En términos algebraicos, las observaciones con posible censura derecha, se representan a través de la siguiente expresión:

$$T_{\mathrm{i}} = \mathrm{mín}\{T_{\mathrm{i}}^{*}, C_{\mathrm{i}}\} \tag{3.8}$$

Donde T_{i} es el tiempo de supervivencia observado y T_{i}^{*} y C_{i}, son los tiempos teóricos de falla y censura (no observables).

Por lo tanto, un dato es censurado cuando el tiempo de censura es menor al tiempo de falla. En términos observables, y de acuerdo a la notación expuesta en 3.1, un cliente es censurado cuando cumple con $\delta_{\mathrm{i}} = 0$.

La importancia de la censura reside en dos aspectos: su relación con la distribución empírica de los datos y su implicancia sobre la función de verosimilitud empleada para estimar los modelos de supervivencia.

Para ilustrar el primer punto se utilizará el ejemplo expuesto por Xu (2001), sobre una muestra de tiempos de remisión de pacientes con leucemia. En efecto, suponga una primera muestra cuyos tiempos de remisión (no censurados) son:

$$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23 \tag{3.9}$$

Con estos datos se pueden calcular las funciones de supervivencia (empíricas) para todas duraciones mediante la fórmula $S_n(t) = \sum_{\mathrm{i}=1}^{n} \frac{I(T_{\mathrm{i}}>t)}{n}$. En particular, se tiene que $S(10) = 8/21$.

Sin embargo, si se asumen datos con algunos tiempos bajo censura derecha (representados con un signo +):

$$6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25, 32+, 32+, 34+, 35+ \tag{3.10}$$

Claramente se puede calcular la función para el tiempo 5 (i.e. $S(5) = 21/21$), pero no se puede calcular la función para el tiempo 7, puesto que no existe certeza de cuánto tiempo sobrevivirá el paciente de observación 6+, lo único que se sabe es que será en un tiempo superior a 6.

De todos modos, se puede calcular $S_n(t)$ asumiendo una muestra sin censura, pero ello no reflejará la función de supervivencia de la población.

Cabe indicar que este inconveniente también tendrá implicancias al momento de evaluar la calidad del pronóstico de los modelos, ya que sería incorrecto comparar curvas predichas (de PI o supervivencia), con esta curva "empírica". No obstante, más adelante se verán formas alternativas para evaluar los modelos.

Por otro lado, respecto del segundo punto, en la literatura se distinguen dos mecanismos de censura: las censuras de tipo 1 y de tipo 2.

La censura de tipo 1 implica que el analista fija de antemano un tiempo de evaluación para cada individuo: si la persona no ha incumplido después de ese tiempo, se declara censurada. Por otro lado, la de tipo 2 declara como censurados a todos los individuos que fallaron después de los r primeros incumplimientos.

La importancia de estudiar los tipos y los mecanismos de censura, radica en sus efectos sobre la función de verosimilitud a estimar. Para el caso de estudio se asumirá censura derecha bajo mecanismo de tipo 1, cuya designación no dependerá de cuánto tiempo esté sobreviviendo el individuo, es decir, se supondrá que el tiempo de censura es independiente del tiempo de falla, aspecto que toma el nombre de *censura no informativa*.

Bajo este esquema, la función de verosimilitud a maximizar (para diferentes tipos de modelos de supervivencia) será:

$$L(\theta) = \prod_{\mathrm{i}=1}^{N} f(t_\mathrm{i}|\theta, Z_\mathrm{i})^{\delta_\mathrm{i}} S(t_\mathrm{i}|\theta, Z_\mathrm{i})^{1-\delta_\mathrm{i}} \tag{3.11}$$

Donde $\delta_\mathrm{i} = 1$ si el individuo incumple al final de la evaluación, θ son los parámetros del modelo y Z_i es el conjunto de variables explicativas del individuo i.

3.2.3. Modelos de supervivencia

Dentro del análisis de supervivencia se distinguen métodos y modelos de estimación, los cuales se describen en la Figura 3.1

Los métodos no paramétricos se caracterizan por estimar una función de supervivencia sin asumir una distribución paramétrica para los tiempos de falla. Entre ellos destaca el método de Kaplan y Meier (1958).

Los modelos paramétricos, son aquellos que requieren asumir una distribución paramétrica para los tiempos de falla y opcionalmente pueden incorporar covariables en el modelo. Entre ellos destacan los *accelerated failure time models* o regresiones AFT (Kalbfleisch y Prentice, 2002).

Mientras que los modelos semi-paramétricos, son aquellos que estiman parámetros por

Figura 3.1: Métodos y modelos de supervivencia

el hecho de incorporar covariables, pero no asumen una distribución paramétrica sobre los tiempos de falla. Entre ellos destacan el modelo de Cox, para el caso de covariables fijas, y el modelo de Cox extendido, para covariables dependientes del tiempo (Cox, 1972).

3.2.4. Método de Kaplan-Meier

El estimador K-M define una función de supervivencia escalonada, obtenida luego de maximizar una forma no paramétrica de la función de verosimilitud en 3.2.2. Se caracteriza por sólo utilizar como información los tiempos de supervivencia y el estatus los clientes al final del periodo, y además por ser capaz de lidiar con la censura derecha.

Los estimadores de Kaplan-Meier para función de supervivencia y la *hazard* instantánea respectivamente son:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{n_j - \mathrm{d}_j}{n_j} \qquad \hat{\lambda}_j = \mathrm{d}_j / n_j \tag{3.12}$$

Donde n_j son los clientes en riesgo justo antes de t_j (es decir, aquellos para los cuales no se sabe si incumplirán o serán censurados en $t \geq t_j$) y d_j son aquellos que incumplen en t_j.

Cabe indicar que la curva de supervivencia de KM es una generalización a la curva de supervivencia empírica, puesto que en ausencia de censura, ambas son equivalentes (Xu, 2001).

Generalmente el método es usado para una primera aproximación a los datos y utilizado como referencia para modelos paramétricos y semi-paramétricos.

Para su cálculo en R se utiliza la función `survfit` de la librería `survival` (T. M. Therneau, 2015).

3.2.5. Regresión AFT paramétrica

Los modelos de tiempos de falla acelerados se basan en la estimación de la siguiente regresión logarítmica:

$$\ln T = \alpha + Z'\beta + \sigma W \tag{3.13}$$

Donde α es el intercepto, β son los coeficientes asociados a las columnas de la matriz de covariables Z (fijas), W es el término de error de la regresión y σ es un parámetro que amplifica el error.

Desarrollando 3.13 se llega a la siguiente expresión (Rodríguez, 2001) (para más detalles diríjase a Anexos):

$$S(t, Z) = S_0(t\mathrm{e}^{-Z'\beta}) \tag{3.14}$$

Donde $S_0(t)$ es la función de supervivencia de un individuo con $Z = 0$ y $\mathrm{e}^{-Z'\beta}$ corresponde al factor de aceleración, el cual acelera o desacelera el recorrido a hacia la falla de acuerdo al valor de las covariables en Z.

En efecto, si se cumple que $\beta > 0$, significa que el proceso de incumplimiento se desacelera a ritmo $\phi = \mathrm{e}^{-\beta} < 1$ y, en consecuencia, la probabilidad de supervivencia es mayor, dado que dicha curva es decreciente en el tiempo.

Por otro lado, la distribución del tiempo de falla T va de la mano con la distribución supuesta para el error. En particular, si se asume una distribución valor extremo para W, se tendrá una distribución Weibull o exponencial para T. Si se asume una distribución normal para W, se tendrá una distribución lognormal para T.

En general, se tendrá que la variable aleatoria $T|Z$ posee una función de distribución $F_W(\gamma \ln(\lambda \mathrm{e}^{-Z'\beta}t))$. Donde $F_W(\tau)$ es la distribución del error W, mientras que $\lambda = \mathrm{e}^{-\alpha}$ y $\sigma = 1/\gamma$ son los parámetros de escala y forma de la variable aleatoria $T|Z = 0$.

Notar que para el individuo nulo (es decir, aquel que tiene $Z = 0$ y que es el necesario para graficar las curvas según 3.14), la función distribución de su tiempo de supervivencia es $F_W(\gamma \ln(\lambda t))$.

Lo expuesto en el párrafo anterior es resumido en la Tabla 3.1 para distribuciones exponencial, Weibull y lognormal.

Cabe señalar que también se pueden definir modelos AFT sin covariables. En dicho contexto sólo se estimará una regresión constante y lo anterior sigue aplicando.

En R, las regresiones AFT paramétricas están implementadas en la rutina `survreg` del paquete `survival`.

Tabla 3.1: **Distribuciones de los tiempos de falla a partir de las distribuciones del error en las regresiones AFT**. Notar que σ no es un parámetro de la distribución valor extremo, sino que es el poderador del error en la regresión AFT. Fuente: elaboración propia.

Distribución del error W	**Distribución de** $T\|Z=0$	**Distribución de** $T\|Z$
V.EXTREMO T1 ($\sigma = 1$)	EXP(λ)	EXP($\lambda e^{-Z'\beta}$)
V.EXTREMO T1 ($\sigma \neq 1$)	WEIBULL(λ,γ)	WEIBULL($\lambda e^{-Z'\beta}$,γ)
NORMAL(0,1)	LOGN($-\ln(\lambda)$,$1/\gamma^2$)	LOGN($-\ln(\lambda e^{-Z'\beta})$,$1/\gamma^2$)

3.2.6. Regresión de Cox

Este tipo de modelos se basa en la siguiente parametrización de la función de riesgo:

$$h(t, Z) = h_0(t) \exp(Z'\beta) \tag{3.15}$$

Donde el primer término se denomina como *baseline hazard function*, la cual se caracteriza por sólo depender del tiempo. Para el caso de los modelos de Cox, la *baseline* es arbitraria, es decir, no se asocia a una distribución paramétrica.

El segundo término, que brinda la información individual de los clientes al modelo, se denomina como *función de riesgo relativa* (Kalbfleisch y Prentice, 2002).

Si se considera la ecuación 3.15, se puede observar que la *baseline hazard* describe el riesgo de los individuos con $Z = 0$, mientras que el término exponencial se interpreta como el riesgo relativo de aquellos con $Z \neq 0$ (Rodríguez, 2001).

Un supuesto fundamental de este modelo es que para dos individuos distintos en la muestra, se tiene que la razón entre sus funciones de riesgo (o *hazard rate*) es invariante en el tiempo. Es decir:

$$HR = \frac{h_0(t) \exp(Z_1'\beta)}{h_0(t) \exp(Z_2'\beta)} = \exp((Z_1 - Z_2)'\beta) = \text{cte} \tag{3.16}$$

Donde Z_1 y Z_2 son las características de dos clientes arbitrarios dentro de la población y β el conjunto de covariables del modelo, las cuales se asumen como únicas para todos los individuos.

Este es el denominado *supuesto de riesgos proporcionales*, el cual se debe testear en los datos para garantizar la viabilidad del ajuste.

De la ecuación 3.15 se puede obtener la función de supervivencia.

$$S(t, x) = S_0(t)^{\exp(Z'\beta)} \tag{3.17}$$

Notar que si $\beta > 0$, la covariable respectiva favorece al incumplimiento. Ello se debe a que ante un cambio unitario en Z se tendrá un riesgo relativo de $\exp(\beta)$, cuyo efecto generará que $S_0(t)^{\exp(\beta)} \leq S_0(t)$, puesto que $S_0(t) \in [0, 1]$.

El hecho de que la *baseline hazard* sea arbitraria reviste inconvenientes para la estimación, puesto que la función de verosimilitud, además de depender de β, dependería de h_0 (i.e. $L(\beta, h_0)$).

Es por ello que Cox (1972) propuso estimar β, maximizando la denominada *verosimilitud parcial*, la cual corresponde a la parte de la verosimilitud que sólo depende de β.

A pesar de que la verosimilitud parcial no tiene interpretación como función de verosimilitud, comparte las mismas propiedades asintóticas, por lo que es factible su uso para muestras grandes (Kalbfleisch y Prentice, 2002).

Sean $t_1 < t_2 < \ldots < t_k$ los distintos tiempos de falla en la muestra y $R(t_j)$ el conjunto de todos los individuos en riesgo en t_j (es decir, aquellos que aún no han incumplido justo antes de t_j). Bajo censura independiente, se tiene que la función de verosimilitud parcial está dada por:

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(Z_j'\beta)}{\sum_{l \in R(t_j)} \exp(Z_l'\beta)} \tag{3.18}$$

La rutina que realiza la estimación del modelo está implementada en la función `coxph` del paquete `survival` de R.

Cuando existen *datos atados*, es decir, cuando hay 2 o más clientes que incumplen en un mismo tiempo de falla t_j (i.e. $\mathrm{d}_j > 1$), existe una aproximación a la verosimilitud propuesta por Breslow (1974).

La verosimilitud de Breslow corresponde a:

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(s_j(t_j)'\beta)}{\left\{\sum_{l \in R(t_j)} \exp(Z_l(t_j)'\beta)\right\}^{\mathrm{d}_j}} \tag{3.19}$$

Donde $s_j(t_j) = \sum_{\mathrm{i}=1}^{\mathrm{d}_j} Z_{j_\mathrm{i}}$ es la suma de las covariables de todos aquellos clientes que incumplen en t_j.

Dicha aproximación se implementa aplicando la opción `ties=‘‘breslow"` en la función `coxph` en R.

Estimando curvas de supervivencia con el modelo de Cox

Dado que en la estimación de β se omitió la influencia de la *baseline*, aún no se pueden definir trayectorias de supervivencia.

Para ello, Klein y Moeschberger (2005) postulan una supervivencia estimada, en base al estimador de la función de riesgo acumulada de Breslow, el cual corresponde a:

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i; \hat{\beta})} \tag{3.20}$$

Donde d_i es el número de incumplimientos en el tiempo t_i, $W(t_i; \hat{\beta}) = \sum_{j \in R(t_i)} \exp(Z'\hat{\beta})$ corresponde a la suma de los *risks scores* (o riesgos relativos) de todos los individuos en riesgo en t_i, y $\hat{\beta}$ el estimador de máxima verosimilitud parcial.

Luego, el estimador de la supervivencia es:

$$\hat{S}_0(t) = -\ln \hat{H}_0(t) \tag{3.21}$$

Dicha estimación se realiza por medio del paquete `survival` de R, utilizando el comando `survfit` (opción `type=‘‘breslow"`) después de haber estimado el modelo de Cox con el comando `coxph`.

Incorporación de estratos

Cuando no se quiere incorporar alguna covariable categórica Z_k en la regresión, se pueden ocupar sus diferentes niveles para definir estratos. La idea es recrear diferentes condiciones base para cada segmento de clientes.

Además de lo anterior, esta estratificación permite lidiar con el incumplimiento del supuesto de proporcionalidad en Z_k, puesto que define una *baseline hazard* específica para cada nivel[1].

El modelo está dado por:

$$\lambda_q(t, x) = \lambda_{0,q}(t) \exp(Z(t)'\beta) \tag{3.22}$$

Donde $q \in 1, \ldots, Q$ son los diferentes estratos.

La función de verosimilitud (parcial) para este modelo, está dada por las contribuciones de cada estrato, las cuales pueden ser calculadas bajo 3.18 o 3.19:

$$L(\beta) = \prod_{q=1}^{Q} L_q(\beta) \tag{3.23}$$

La estimación se realiza a través de la opción `strata` dentro de la función `coxph` del paquete `survival` de R.

Teniendo ya estimado β, para definir curvas de supervivencia por segmento, se puede aplicar el método de estimación de la *baseline* antes descrito.

[1]Para el caso de variables continuas, se recomienda su discretización a intervalos.

Supuesto de riesgos proporcionales

Los modelos de Cox (estándar) trabajan bajo el supuesto de que la razón entre los efectos de cada covariable, son fijos. No obstante, esto hay que testearlo.

Para ello Grambsch y Therneau (1994) propusieron un test χ^2 para la proporcionalidad, basado en residuos de Schoenfeld (ver Anexo 3), el cual busca verificar si los efectos de las covariables son constantes en el tiempo.

Si estos efectos no son constantes, significa que las *hazard rates* entre dos individuos cualquiera en la muestra, cambiarán en el tiempo, rompiendo con la proporcionalidad.

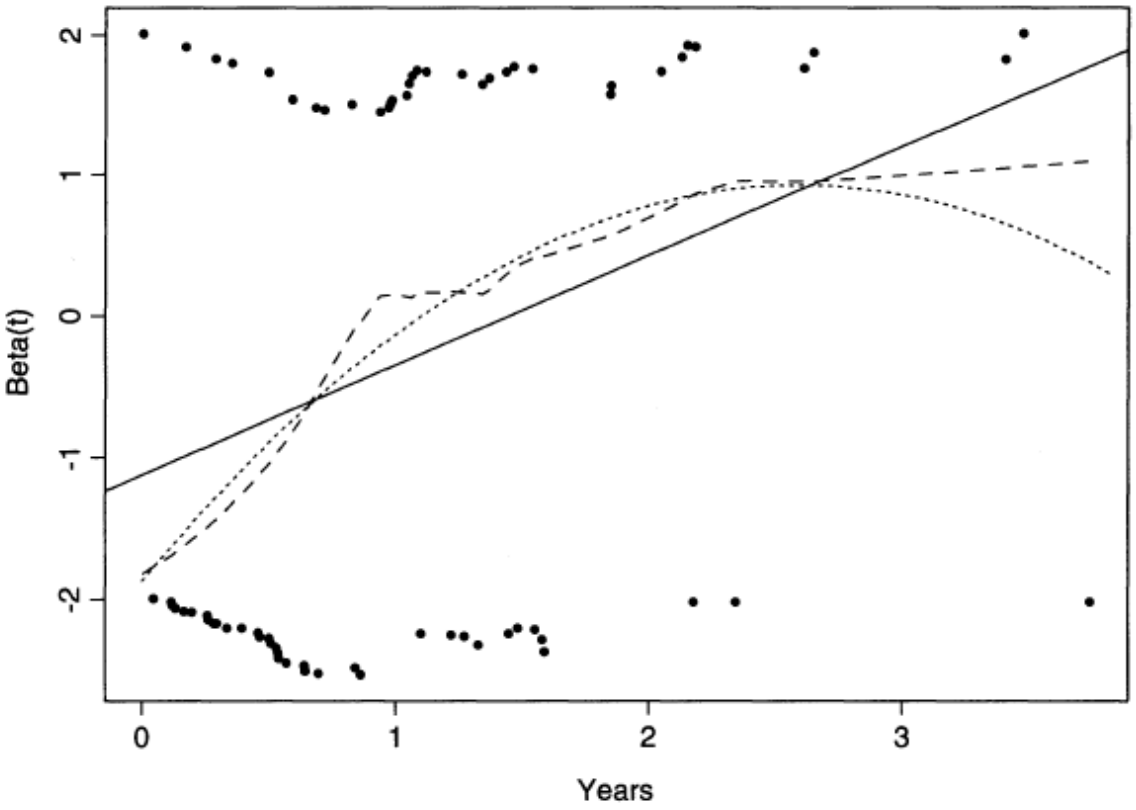


Figura 3.2: **Residuos de Schoenfeld en función al tiempo**. Este gráfico aplica diferentes ajustes sobre los residuos: un ajuste lineal, un ajuste cuadrático y un ajuste suave (*lowess*). Si la pendiente del ajuste lineal es 0, entonces no se rechazará la hipótesis de riesgos proporcionales para la variable evaluada. Fuente: T. M. Therneau y Grambsch (2000)

Sin embargo, T. M. Therneau y Grambsch (2000) señalan que el test puede fallar para muestras grandes, por lo que para esos casos, recomiendan evaluar métodos gráficos. En específico, proponen ver el diagrama de dispersión entre los residuos de Schoenfeld y el tiempo, tal como se ilustra en la Figura 12.1. En caso de que un ajuste suave (tipo *lowess*) tenga una pendiente aproximadamente 0, significa que no se rechazará el supuesto de riesgos proporcionales.

Tanto el test como los diagramas de dispersión están implementados en la función `cox.zph`

en R.

3.2.7. Regresión de Cox extendida

El hecho de considerar sólamente el valor inicial de las covariables para cada cliente, hace pensar que se está desaprovechando una gran cantidad de datos para la estimación.

Además, si dichas covariables cambian en el tiempo, ello implica automáticamente la violación del supuesto de riesgos proporcionales anteriormente descrito.

La razón de ello es que el producto entre coeficientes dependientes del tiempo y las variables iniciales, pueden ser vistos como el producto entre un coeficiente fijo y una covariable dependiente del tiempo (Collett, 2015)[2]. Es decir:

$$\beta(t) \cdot X = \beta \cdot X(t) \tag{3.24}$$

En este caso la función de verosimilitud parcial se generaliza incorporando la dimensión temporal en las covariables[3].

$$L(\beta) = \prod_{j=1}^{k} \frac{\exp(Z_j(t_j)'\beta)}{\sum_{l \in R(t_j)} \exp(Z_l(t_l)'\beta)} \tag{3.25}$$

Sobre ella también se puede aplicar la aproximación expuesta en 3.19.

Otra diferencia entre estos modelos y aquellos con covariables fijas, corresponde a la codificación de su base de datos: en los modelos con covariables fijas, se utiliza una base agregada por cliente, mientras que en los modelos con variables dependientes, se utiliza un panel de datos individuo-tiempo con ciertas modificaciones que serán expuestas en el Capítulo 4.

3.2.8. Evaluación de modelos

Dentro de los métodos de evaluación en regresiones AFT y modelos de Cox, existen aquellos basados en criterios gráficos y otros basados en métricas, los cuales se resumen en la Tabla 3.2 según el tipo de modelo.

Residuos de Cox-Snell

Para medir la bondad de ajuste a nivel general del modelo, se ocuparán los residuos de Cox-Snell.

[2]Recordar que la hipótesis alternativa del test de riesgos proporcionales corresponde a la dependencia temporal del coeficiente β.

[3]Para detalles más teóricos se recomienda leer los Capítulos 4 y 6 de Kalbfleisch y Prentice (2002)

Tabla 3.2: **Métodos de evaluación para regresiones AFT y modelos de Cox.** Existen criterios gráficos y criterios en base a métricas. Los criterios gráficos evalúan la bondad de ajuste para una misma clase de modelos (AFT ó Cox), dichos criterios no son comparables entre diferentes clases. El AIC y el pseudo R^2 son métricas de bondad de ajuste, mientras que el C-index y el iAUC son métricas de discriminación. Las métricas AIC no son comparables entre modelos de Cox y AFT, ya que la primera se basa en la verosimilitud y la segunda, en la verosimilitud parcial. Las métricas AIC y los C-index sí son comparables entre los dos tipos de modelos de Cox. Fuente: Elaboración propia.

Método de evaluación	**Regresión AFT**	**Regresión de Cox**	**Regresión de Cox extendida**
Criterios gráficos	**Residuos de Cox-Snell** (Collett, 2015) **Gráficos de diagnóstico** (Klein y Moeschberger, 2005)	**Residuos de Cox-Snell** (Collett, 2015; Klein y Moeschberger, 2005; Therneau y Grambsch, 2000)	-
Métricas	-	**AIC** **C-index** (Harrell et al, 1996) **Pseudo R^2** (Nagelkerke, 1991) **iAUC** (Heagerty y Zheng, 2005; Saha y Heagerty, 2010)	**AIC** **C-index** (Harrell et al, 1996)

Sean $\{t_{\mathrm{i}}\}_{\mathrm{i}=1}^{n}$ los diferentes tiempos de falla para los n individuos de la muestra, se definen los residuos de Cox-Snell para los modelos AFT paramétricos, como:

$$r_j = \hat{H}_{\mathrm{i}}(t_j) = -\ln \hat{S}_{\mathrm{i}}(t_{\mathrm{i}}), \mathrm{i} = 1, \ldots, n \tag{3.26}$$

Donde $\hat{H}_{\mathrm{i}}(t_j)$ y $\hat{S}_{\mathrm{i}}(t_{\mathrm{i}})$ son las funciones de riesgo acumulada y de supervivencia estimadas para el individuo de covariables Z_{i}, a partir de una distribución supuesta para el tiempo de falla.

Los residuos de Cox-Snell para los modelos de Cox, se definen como:

$$r_j = \exp(Z_{\mathrm{i}}'\hat{\beta})\hat{H}_0(t_{\mathrm{i}}), \mathrm{i} = 1, \ldots, n \tag{3.27}$$

Donde $\hat{H}_0(t_{\mathrm{i}})$ es la función de riesgo acumulada de nivel base estimada en base a un método no-paramétrico (por ejemplo, Breslow).

Si la muestra de los residuos distribuye como una exponencial de parámetro 1, significa que el modelo es correcto.

Para ello se debe aplicar un ajuste de Cox o AFT (según corresponda), utilizando como variable temporal los residuos de Cox-Snell y como variable de evento el estatus de incumplimiento. Posteriormente se grafica la función de riesgo acumulada versus los residuos y se verifica si se acerca a la identidad.

Gráficos de diagnóstico

Este método es sólo aplicable a los modelos paramétricos. Su uso es para determinar si una distribución es adecuada para los tiempos de falla. El elemento basal es la función de riesgo acumulada (poblacional) de la distribución asumida.

El procedimiento consiste en aplicar transformaciones sobre dicha función, de modo de obtener una relación lineal con el tiempo. Esto dará una identidad que será siempre válida para la distribución asumida.

Por ejemplo, para la distribución Weibull se tiene que $H(t) = \lambda t^{\gamma}$, por lo tanto, para tener una relación lineal con el tiempo, se debe aplicar logaritmo, obteniendo la siguiente identidad:

$$\ln H(t) = \ln \lambda + \gamma \ln t \tag{3.28}$$

Posteriormente, se debe verificar si al evaluar los datos se sigue cumpliendo la identidad. Para ello, se utiliza el estimador de Kaplan-Meier y se verfica si las curvas son lineales. Para el caso exponencial, éstas deben ser además de pendiente 1.

Criterio de información de Aikake

Esta métrica permite evaluar el modelo en función a su verosimilitud y su complejidad (la métrica castiga la presencia de covariables). Su formulación se describe como:

$$AIC = 2k - 2\ln L(M) \tag{3.29}$$

Donde k corresponde al número de covariables y $L(M)$ es la verosimilitud del modelo M. Cuan menor es su valor, mejor es el modelo.

C-index de Harrell

El C-index, se define como la proporción de todos los pares de individuos comparables, cuyas predicciones y tiempos observados de falla son concordantes (Harrell, Lee, y Mark, 1996).

Se considerarán pares no comparables aquellos que:

- Fallan al mismo tiempo.
- Tienen iguales tiempos de falla.
- Uno de ellos es censurado en un tiempo inferior al momento de falla del otro (no hay certeza si falla antes o después del segundo).

Sobre los pares restantes (comparables) se define un par concordante cuando la observación con un tiempo de falla menor, tiene mayor *risk score*; lo contrario ocurre con los pares discordantes. Un par es atado (*tied*) cuando presentan predictores equivalentes.

Finalmente la fórmula del C-index está dada por:

$$\text{C-index} = \frac{\text{concordantes} + \frac{\text{atados}}{2}}{\text{concordantes} + \text{discordantes} + \text{atados}} \tag{3.30}$$

Esta métrica se reporta en los resultados de `coxph` y está disponible para modelos de Cox estándar y extendidos

Pseudos R^2 de Cox-Snell y Nagelkerke

El pseudo-R^2 de Cox-Snell está en el contraste entre la verosimilitud de un modelo con covariables versus un modelo nulo. Éste está definido por:

$$R^2_{\text{C-S}} = 1 - \left(\frac{L(M_{\text{intercept}})}{L(M_{\text{full}})}\right)^{2/N} \tag{3.31}$$

Donde $L(M_{\text{intercept}})$ corresponde a la función de verosimilitud del modelo sin covariables (Kaplan-Meier) y $L(M_{\text{full}})$ es el valor para el modelo con covariables. N es el total de las observaciones.

Notar que su valor está acotado entre 0 y $1 - (L(M_{\text{intercept}}))^{2/N}$.

Cabe señalar, que debido a lo anterior Nagelkerke (1991) redefinió el R^2 de Cox-Snell, de modo que estuviera acotado entre valores de 0 y 1. En efecto, el R^2 de Nagelkerke está dado por:

$$R^2_{\text{N}} = \frac{R^2_{\text{C-S}}}{\text{máx}\, R^2_{\text{C-S}}} \tag{3.32}$$

Donde el máximo R^2 de Cox-Snell está dado por $1 - (L(M_{\text{intercept}}))^{2/N}$.

Esta métrica se reporta en los resultados de `coxph` y está disponible para modelos de Cox estándar y extendidos.

Integral de los AUC dependientes del tiempo

El iAUC (Heagerty y Zheng, 2005; Saha y Heagerty, 2010), equivale a un promedio ponderado de las áreas bajo las curvas ROC (los AUC) en el tiempo.

Su fórmula está dada por (para detalles en su derivación, diríjase al Anexo 3):

$$\text{iAUC} = C^\tau = \int_0^\tau \text{AUC}(t) \cdot \omega^\tau(t) \text{d}t \tag{3.33}$$

Donde $\omega^\tau(t) = 2f(t) \cdot S(t)/W^\tau$ y $W^\tau = \int_0^\tau 2f(t) \cdot S(t)/W^\tau \text{d}t = 1 - S^2(t)$.

Es una medida de concordancia al igual que el C-index, puesto que verifica si los riesgos relativos de cada par de clientes son concordantes con sus tiempos de falla (i.e. un cliente con mayor riesgo relativo debe tener un tiempo de falla menor que su par). En la práctica, presenta diferencias (pequeñas) con el C-index.

La métrica está implementada en la función `risksetAUC` de librería `riskSetROC` de R, la cual sólo se aplica para modelos de Cox estándar.

3.2.9. Resumen de los modelos

Tabla 3.3: **Resumen de los modelos de supervivencia**. Notar que los modelos de Cox extendidos requieren de mayor procesamiento, pero para este trabajo de memoria los tiempos de ejecución fueron similares entre todos los modelos. Fuente: Elaboración propia.

Dimensión	**Regresión AFT**	**Regresión de Cox estándar**	**Regresión de Cox extendida**
Covariables	Fijas.	Fijas.	Covariables dependientes del tiempo.
Pronóstico para tiempos arbitrarios	Sí	No	No
Heterogeneidad	Individual (y grupal, si se toma individuo promedio).	Individual (y grupal, si se toma individuo promedio).	Individual (y grupal, si se toma individuo promedio).
Distribución del tiempo de falla	Se debe asumir.	No es necesario asumir.	No es necesario asumir.
Supuestos del modelo	Aceleración/desaceleración del tiempo de falla.	Riesgos proporcionales.	Covariables externas (ver Capítulo 9).
Bondad de ajuste	Gráficos de diagnóstico y residuos de Cox-Snell.	Residuos de Cox-Snell, AIC y Pseudo R^2	AIC.
Discriminación	-	C-index, iAUC.	C-index.
Función de la PI estimada	Explícita del tiempo.	No es explícita del tiempo.	No es explícita del tiempo.
Requerimientos de información (tanto para calibrar, como para pronosticar)	Información al inicio del seguimiento.	Información al inicio del seguimiento.	Historia del cliente.

Capítulo 4

Metodología

En el presente capítulo se describirán todas las actividades adscritas al estudio, las cuales fueron divididas en 4 fases similares a los expuestos en el enfoque KDD (Fayyad, Piatetsky-Shapiro, y Smyth, 1996).

La primera de ellas corresponde al pre-procesamiento y transformación de los datos, en la cual se aplicarán los filtros de datos y se construirán las bases finales, las que se emplearán tanto para fines exploratorios como para entrenar los modelos.

La segunda corresponde a la exploración de datos, la cual pretende buscar patrones y sentar las hipótesis en torno a los efectos de variables sobre la probabilidad de incumplimiento.

La tercera fase es el procesamiento de los datos, instancia en la cual se entrenarán tres clases de modelos (AFT, Cox estándar y Cox extendido) y se evaluarán diferentes combinaciones de covariables para cada uno de ellos. Además, se verificarán sus poderes discriminativos y bondades de ajuste, por medio de métodos gráficos y métricas estadísticas.

La cuarta fase corresponde a la comparación entre las distintas clases de modelos, la cual se realizará a través de argumentos preferentemente operacionales.

La Figura 4.1 resume las fases con los respectivos pasos a seguir.

4.1. Pre-procesamiento y transformación

4.1.1. Filtro de datos

Para llevar a cabo el filtro de datos se aplicó un código programado en lenguaje R, el cual tomó como insumo un panel de datos provisto por la empresa en formato *.txt*, el cual cuenta con múltiples observaciones de 10.000 clientes de su cartera de consumo. En particular, para cada cliente se tiene el tiempo (calendario) de registro, su estatus de incumplimiento y otras variables explicativas.

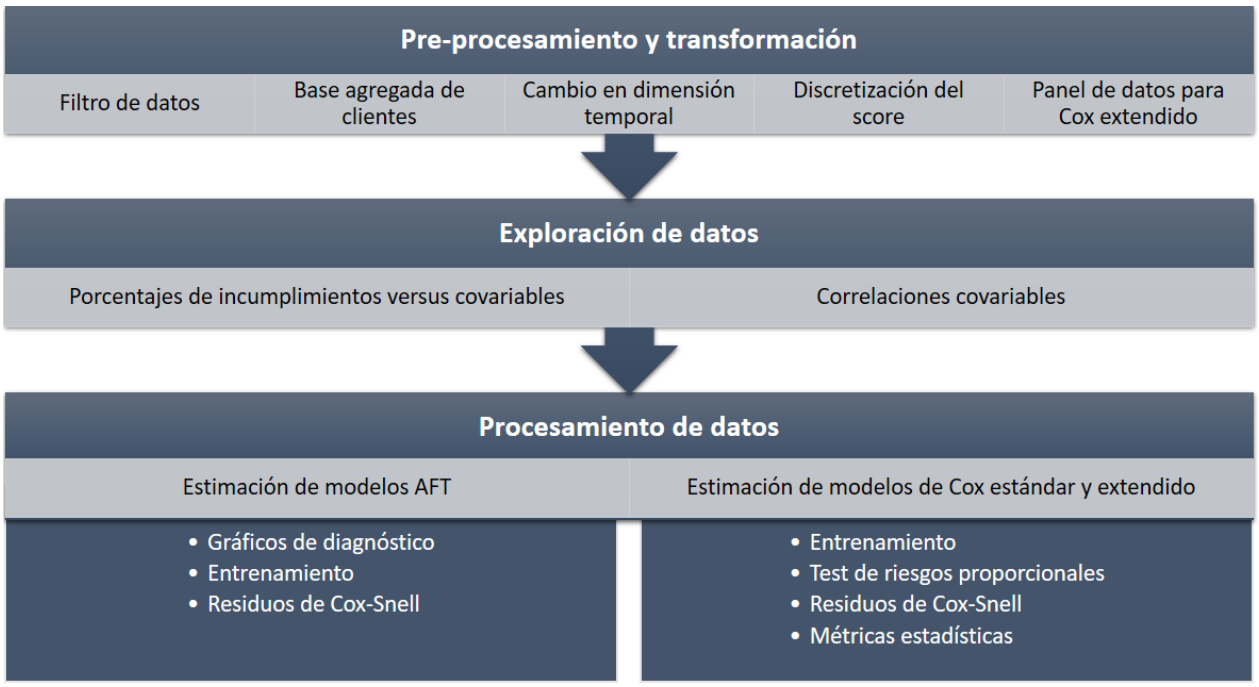


Figura 4.1: Fases metodológicas y pasos a seguir

Lo que hace este código es aplicar las siguientes reglas:

- Para cada cliente, eliminar todas las observaciones después de haber alcanzado el estatus de incumplimiento.
- Para todos los clientes sin registro de la variable explicativa score, reparar los datos con el registro más cercano en el tiempo.
- Eliminar todos los clientes sin información de sus variables explicativas a lo largo del estudio (431 clientes).
- Eliminar a todos los clientes bajo incumplimiento al inicio del estudio (censura izquierda) (35 clientes dentro de los 431).

Tras aplicar el código se mantuvo la información de 9569 clientes.

4.1.2. Base de agregada de clientes

Para esta tarea se aplicó otro script de R sobre el panel de clientes filtrado, en él se agrega la información para cada cliente, obteniendo su tiempo de vida (duración), su estatus de incumplimiento al final del seguimiento y múltiples agregaciones de sus variables explicativas (variables al inicio, variables promedio, variables máximas, mínimas y finales).

Su construcción se debe a que las regresiones AFT y el modelo de Cox estándar requieren de covariables fijas, por lo que sólo admite una base donde cada fila sea un cliente.

4.1.3. Cambio de dimensión temporal

Cabe mencionar que uno de los pasos en el apartado anterior fue obtener el tiempo de vida del cliente a partir de sus múltiples observaciones. En efecto, lo que se hizo ahí fue cambiar el tiempo desde una dimensión de fecha calendario a una dimensión de duración, tal como se realizó en los trabajos de Bogren (2015) y Man (2014).

A modo de ejemplo, en las Figuras 4.2 y 4.3 se muestran dos diagramas de eventos bajo diferentes dimensiones temporales para 6 clientes de la cartera.

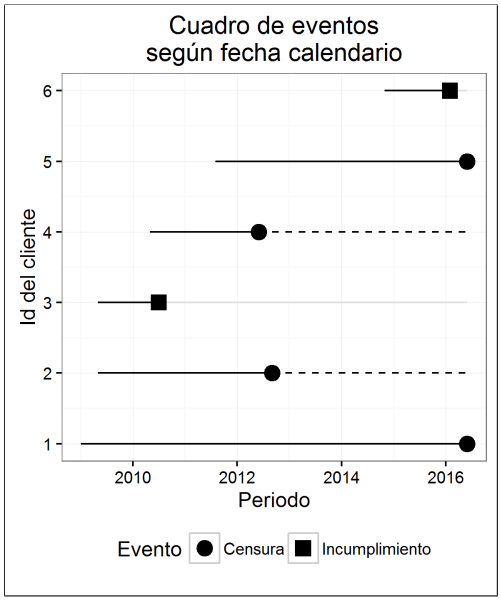


Figura 4.2: Diagrama de eventos para 6 clientes de la cartera de consumo según fecha calendario.

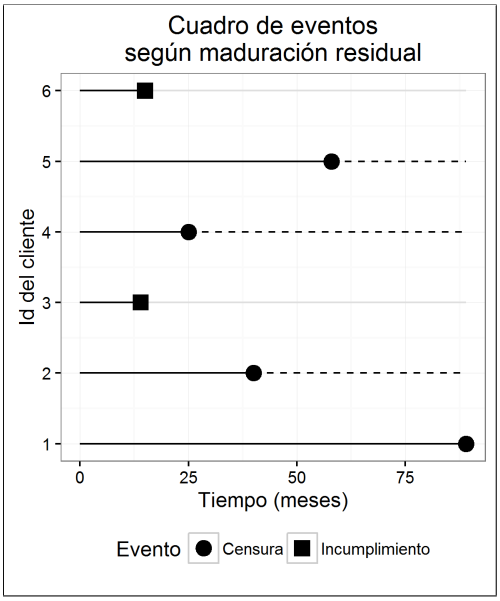


Figura 4.3: Diagrama de eventos para 6 clientes de la cartera de consumo según maduración.

4.1.4. Discretización del score

Como uno de los objetivos es describir curvas de PI para distintos segmentos de clientes, se optó por segmentarlos según tramos de score, a través de un tercer código en R.

La idea de usar tramos de score se debe a la costumbre de la empresa por utilizar esta métrica como insumo para otros modelos internos.

Notar que los tramos de score podrían entrar en juego mediante dos posibles formas: aplicar múltiples modelos para cada submuestra inducida por la segmentación, o bien, utilizar los segmentos como una variable categórica en los modelos de regresión. Dicha discusión se dará más adelante.

La segmentación se realizó a través de K-means sobre el score inicial, para modelos de Cox y AFT con covariables al inicio; y sobre el score promedio, para regresiones AFT con

covariables al promedio y modelos de Cox con covariables dependientes del tiempo.

Ello permitió elaborar los rankings expuestos en las Tablas 4.1 y 4.2, con los cuales se pudieron asignar los grupos a los clientes de acuerdo a su score inicial, score promedio, o a su score en el tiempo, según el modelo correspondiente.

Tabla 4.1: **Segmentos de score inicial**. Muestra los rangos asociados a cada segmento de score inicial, los cuales fueron obtenidos a partir del método K-Means de una variable sobre el score inicial de cada cliente.

Segmento	Intervalo de score inicial	Tamaño (N clientes)	Riesgo
1	(73,454]	1698	Alto
2	(454,616]	1930	
3	(616,745]	1242	
4	(745,859]	1531	
5	(859,957]	828	
6	(957,1052]	2075	
7	(1052,1198]	265	Bajo

Tabla 4.2: **Segmentos de score promedio**. Muestra los rangos asociados a cada segmento de score promedio, los cuales fueron obtenidos a partir del método K-Means de una variable sobre el score promedio de cada cliente. Cabe indicar que esta tabla se utilizará tanto para ajustar regresiones AFT evaluadas en covariables promedio, como para modelos de Cox extendidos. En este último caso, como el score cambia a lo largo del tiempo, se asignarán los segmentos de forma dinámica, es decir, en cada momento del tiempo se verificará a qué rango de la tabla pertenece el score.

Segmento	Intervalo de score promedio	Tamaño (N clientes)	Riesgo
1	(73,443]	2784	Alto
2	(443,646]	1558	
3	(646,793]	1544	
4	(793,886]	1161	
5	(886,981]	443	
6	(981,1082]	1838	
7	(1082,1198]	241	Bajo

4.1.5. Panel de datos ajustado para modelos de Cox extendidos

Anteriormente se había elaborado una base agregada de clientes con el objeto de entrenar modelos con covariables fijas. No obstante, para estimar modelos con covariables dependientes del tiempo se requieren de múltiples observaciones para cada cliente. Por lo tanto, se debe trabajar sobre el panel inicial y no sobre la base agregada.

Sin embargo, antes de emplear el panel inicial, se le debe cambiar de formato, de modo que sea compatible con la rutina `coxph` del paquete `survival` de R, empleada para estimar el modelo.

Para ello se aplicó la función `tmerge` (de la misma librería), la cual sustancialmente transforma los tiempos puntuales de registro en una notación de intervalos.

Para explicar el funcionamiento se utilizará un ejemplo dentro del documento de T. Therneau, Crowson, y Atkinson (2016):

Considere dos bases de datos: una con información inicial de cada cliente más su tiempo de supervivencia T_n, y otra que contiene sus mediciones a lo largo del tiempo. Supóngase que se quiere "pegar" la información de la segunda base sobre la primera, pero dividiendo consecutivamente el intervalo $[0, T_n]$ en múltiples sub-intervalos $[0, T_1], (T_1, T_2], \ldots (T_{n-1}, T_n]$, a medida que se van incorporando nuevas mediciones.

En este caso, cada T_j corresponde al tiempo durante el cual ocurre un incumplimiento, o bien, cambia de valor alguna de las covariables. En efecto, el funcionamiento es similar a la acción de colocar múltiples cartas entremedio de una baraja.

4.2. Exploración de datos

La exploración de datos se realizará por medio de gráficos dinámicos en Microsoft Excel y además se determinará la correlación de las covariables en R. En particular se ilustraron los porcentajes de incumplimiento para diferentes valores en las covariables y se elaboraron gráficos de dispersión tiempo-covariables. Con ello se postularon hipótesis para ver si se cumplen luego de estimar los modelos.

4.3. Procesamiento de datos

El procesamiento de datos estará cronológicamente separado en 3 etapas: estimación de los modelos AFT paramétricos, estimación de los modelos de Cox estándar y estimación de los modelos de Cox extendidos.

4.3.1. Estimación de modelos AFT paramétricos

En esta primera etapa, se aplicará un script de R que realizará lo siguiente: elaborar gráficos de diagnóstico, entrenar las regresiones AFT y graficar los residuos de Cox-Snell.

La elaboración de gráficos de diagnóstico (tal como se estipula en el marco teórico), busca generar una idea en torno a la distribución más adecuada para los tiempos de falla.

Para ello se aplicará la función `survfit` (librería `survival` de R) con el fin de obtener la función de riesgo acumulada de Kaplan-Meier. Luego, con el comando `ggplot` (librería `ggplot2` de R) se graficará dicha función (o alguna transformación de ella, según sea el caso) con respecto al tiempo para ver si la relación entre ambos es lineal.

Para estimar regresiones AFT se utilizará el comando (librería `survival`), el cual entregará como output los coeficientes de las regresiones y otros resultados útiles.

Un ejemplo del funcionamiento del comando es el siguiente:

```
regAFT <- survreg( Surv( time, incump ) ~ antIni  + moraIni + renegIni +
                              factor( scoreIniCat ), data = df, dist = ‘‘weibull" )
```

Lo que realiza este comando es crear un objeto de clase `survreg`, el cual contiene resultados de una regresión AFT Weibull (coeficientes, predictores lineales para cada cliente, verosimilitud, etc.).

Esta expresión indica que se utilizaron como covariables: la antigüedad inicial de la cuenta corriente, los días de mora inicial, el estatus de renegociación inicial y múltiples variables dummies obtenidas luego de aplicar el operador `factor` sobre el score inicial categórico. Además, como fuente de datos, se utilizó la base `df`, la cual contiene la información de los clientes.

Este tipo de regresiones se aplicó para otras distribuciones (exponencial y lognormal) y se consideraron diferentes combinaciones de covariables (incorporando algunas y sustrayendo otras).

Para obtener los residuos de Cox-Snell, primero se debieron obtener los residuos estandarizados (Collett, 2015), los cuales, al igual que en una regresión lineal, corresponden a la diferencia entre el entre la variable dependiente y su predictor, pero estandarizadas por el predictor del término que acompaña al error en la regresión AFT (i.e. $\hat{\sigma}$). En este caso la variable dependiente es el logaritmo de los tiempos de falla y predictor es la suma ponderada de las características. Estos residuos se obtienen por medio de la función `predict` (`type = ‘‘working"`) sobre el objeto `survreg`.

Luego, los residuos de Cox-Snell se obtienen aplicando una transformación sobre los residuos estandarizados, específica a cada distribución asumida.

Finalmente, para evaluar la bondad de ajuste del modelo, se debe graficar la función de riesgo acumulada de los residuos de Cox-Snell. Si los puntos se acercan a la identidad (i.e. que los residuos distribuyan exponencial de tasa 1), significa que el modelo es bueno.

4.3.2. Estimación de los modelos de Cox estándar y extendido

En este caso se procedió a estimar inmediatamente las regresiones de Cox estándar y extendidas, por medio de las siguientes expresiones en R, respectivamente:

```
regCox <- coxph( Surv( time, incump ) ~ antIni + moraIni + renegIni +
                        factor( scoreIniCat ), data = df, ties = ‘‘breslow" )

regCoxExt <- coxph( Surv( tstart, tstop, inc ) ~ ant  + mora + reneg +
                factor( scoreCat ), data = df_tdc_fscore, ties = ‘‘breslow" )
```

En el primer caso, al igual que en el modelo AFT, se define un objeto `coxph` evaluado con las covariables al inicio y aplicado sobre la base agregada de clientes. La opción `ties` indica qué ajuste a la verosimilitud parcial se debe hacer para lidiar con los datos atados; en este caso se aplicará la aproximación de Breslow. Notar que en este caso time hace referencia al tiempo de vida del cliente, mientras que incump lo hace para el estatus de incumplimiento al final del seguimiento.

En el segundo caso se observan diferencias, puesto que en el brazo izquierdo de la fórmula se ocupan los parámetros `tstart` y `tstop`, los cuales definen múltiples periodos de tiempo en formato de intervalo. Además, `inc` indica el estatus de incumplimiento en diferentes momentos del tiempo.

En el brazo derecho de la fórmula las variables `ant`, `mora`, `reneg` y `scoreCat`, hacen referencia a la antigüedad, mora, renegociación y score categórico en diferentes momentos del tiempo. Mientras que `data_tdc_fscore`, es la base de datos ad-hoc para los modelos de Cox extendidos, la cual considera el score categórico como covariable dependiente del tiempo.

Luego de la estimación, se debe verificar el cumplimiento del supuesto de riesgos proporcionales en cada una de las variables del modelo de Cox estándar. Esto se hace a través de la función `cox.zph` sobre el objeto de clase `coxph`, la cual dará entregará el resultado del test χ^2 de (Grambsch y Therneau, 1994). Además, se aplicará la función `plot` sobre el objeto `cox.zph` para graficar los residuos de Schoenfeld en el tiempo, lo cual servirá para complementar u objetar los resultados del test (recordar que el test tiende a rechazar la hipótesis para muestras grandes).

Posteriormente, se evaluará la bondad de ajuste del modelo de Cox estándar por medio de los residuos de Cox-Snell, los cuales se calculan, para cada cliente, a partir de la diferencia entre el estatus de incumplimiento y los residuos de martingala[1].

Éstos últimos se obtienen luego de aplicar la función `predict` (`option = ‘‘martingale"`) sobre el objeto `coxph`. Luego de ello, se grafica el riesgo acumulado versus los mismos residuos y se verifica si se ajusta a la identidad.

Finalmente, se obtendrán los indicadores de ajuste estadístico según la clase de modelo de Cox a evaluar: AIC, C-index, iAUC y pseudo R^2 para el modelo de Cox estándar; y AIC y C-index para el modelo de Cox extendido.

[1]Esto es sólo válido para modelos de Cox con covariables fijas

Capítulo 5

Exploración de datos

5.1. Datos

La entidad bajo estudio facilitó una base de datos con la información de 10.000 clientes de su cartera de consumo. El panel de datos considera, para cada cliente, distintas variables en diferentes momentos del tiempo. En total el panel contempla 548.996 registros entre Diciembre de 2008 hasta Mayo de 2016. La variables del panel se describen en la Tabla 5.1.

Tabla 5.1: Variables del panel inicial de datos.

Variable	**Descripción**
Desempeño P12M	Variable que vale 1 si el cliente es considerado Malo en los 12 meses próximos, según los criterios internos de la entidad bancaria[1].
Es Malo	Vale 1 si el cliente es considerado malo en la fecha actual.
Score	Puntaje del cliente (en términos de su capacidad crediticia) extraido del modelo de credit score subyacente.
Renegociado	Variable dummy que indica si en la fecha actual si el cliente ha renegociado un crédito, o no.
Ant CCT	Antigüedad de la cuenta corriente del cliente.
Fecha	Fecha calendario actual.
Rut	String anonimizado que hace referencia única al cliente.

Después de haber realizado un pre-procesamiento y transformación de los datos, se obtuvo una base agregada que contiene la información de cada cliente (ver Capítulo 4). Las principales variables de esta base agregada se muestran en la Tabla 5.2.

Tabla 5.2: Variables principales de la base agregada de clientes

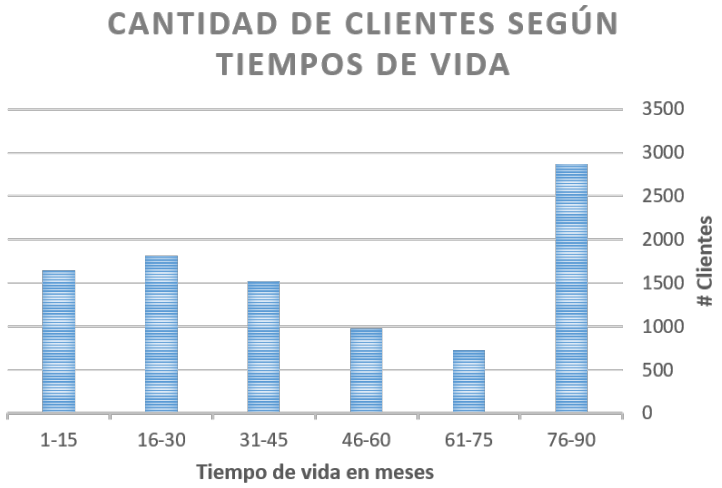
Variable	**Referencia en código R**	**Descripción**
Rut	`rut`	String anonimizado que hace referencia única al cliente
Fecha inicial	`f_ini`	Fecha de inicio de seguimiento
Fecha final	`f_fin`	Fecha final de seguimiento
Tiempo de vida, de falla, o supervivencia	`time`	Diferencia en meses entre la fecha de inicio y fin de seguimiento
Incumplimiento	`incump`	Indicatriz que vale 1 cuando el cliente es considerado "Malo"(i.e. bajo incumplimiento) en la fecha final de seguimiento. Si este valor vale 0, significa que el cliente presenta censura derecha.
Score inicial	`scoreIni`	Puntaje al inicio del seguimiento.
Score promedio	`scoreProm`	Puntaje promedio a lo largo del seguimiento.
Score categórico inicial	`scoreIniCat_k1_7`	Segmento asociado al cliente según su valor de score inicial. Los segmentos fueron definidos a través del método k-means de una variable con un límite de 7 clústeres.
Score categórico promedio	`scorePromCat_k1_7`	Segmento asociado al cliente según su valor de score promedio. Los segmentos fueron definidos a través del método k-means de una variable con un límite de 7 clústeres.
Mora inicial	`moraIni`	Días de mora al inicio del seguimiento.
Mora promedio	`moraProm`	Días de mora al final del seguimiento.
Antigüedad inicial	`antIni`	Antigüedad de la cuenta corriente del cliente al inicio del seguimiento.
Antigüedad promedio	`antProm`	Antigüedad promedio de la cuenta corriente del cliente.
Renegociación inicial	`renegIni`	Estatus de renegociación del cliente al inicio del seguimiento. Toma el valor 1 si ha presentado renegociación, 0 si no.
Renegociación en la vida del cliente	`renegDummy`	Estatus de renegociación del cliente en algún momento del seguimiento. Toma el valor 1 si ha presentado renegociación, 0 si no.

La Tabla 5.3 muestra las estadísticas descriptivas del tiempo de vida, mientras que la Figura 5.1 muestra el histograma de los tiempos de vida de los clientes (incluyendo clientes censurados y bajo incumplimiento).

Tabla 5.3: Estadísticas descriptivas de los tiempos de vida.

Mínimo	Percentil 25 %	Mediana	Promedio	Percentil 75 %	Máximo
1.00	21.00	43.00	48.33	81.00	89.00

Figura 5.1: Histograma de los tiempos de vida.



Del histograma se puede observar que la mayoría de los clientes tiene un tiempo de vida entre 76 y 90 meses. Sin embargo, excluyendo ese tramo se observa una disminución desde el segundo al cuarto tramo.

Lo anterior podría ser contra-intuitivo, puesto que se esperaría una caída monótona en la cantidad de clientes a medida que avanza en el tiempo.

Sin embargo, este hecho se atribuye a que, en la cartera de consumo, un mismo cliente puede tener múltiples instrumentos a su haber. Se conjetura que, de evaluar un solo instrumento, podría darse un decrecimiento en los clientes a lo largo del tiempo[2].

De todos modos, de acuerdo con la opinión de los profesionales de la entidad, 90 meses sin incumplir sigue siendo un tiempo razonable.

5.2. Variables, incumplimiento y tiempos de falla

A continuación se verá la relación de la antigüedad, los días de mora, la renegociación y el score fueron variables con el incumplimiento y el tiempo de falla.

[2]Para verificar lo anterior, es necesario disponer de datos operacionales, los cuales entregan el detalle del producto. Dado que en este estudio sólo se cuenta con la información de los clientes, no se puede hacer el análisis anterior, sin embargo, quedará como propuesto.

Figura 5.2: Porcentajes de incumplimiento y tiempos de falla según antigüedad inicial y promedio.

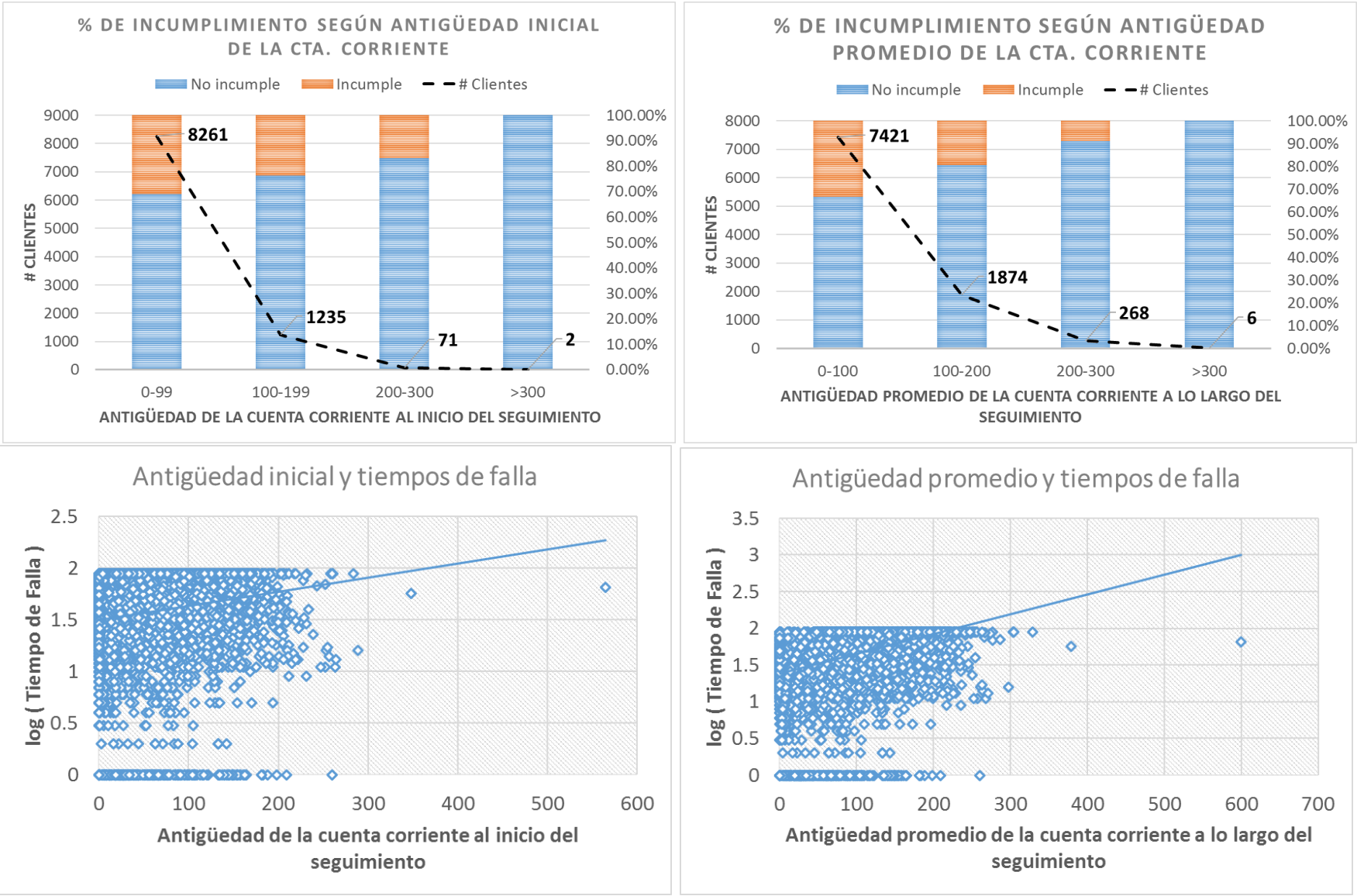
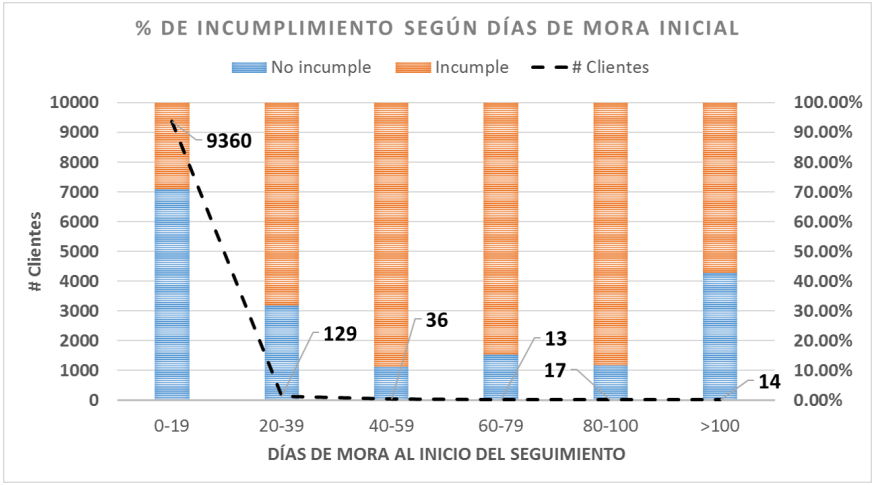
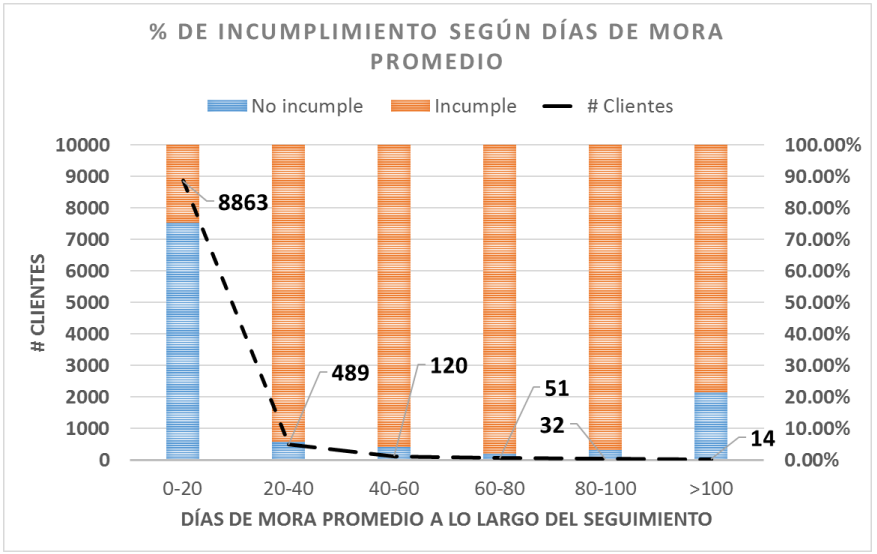
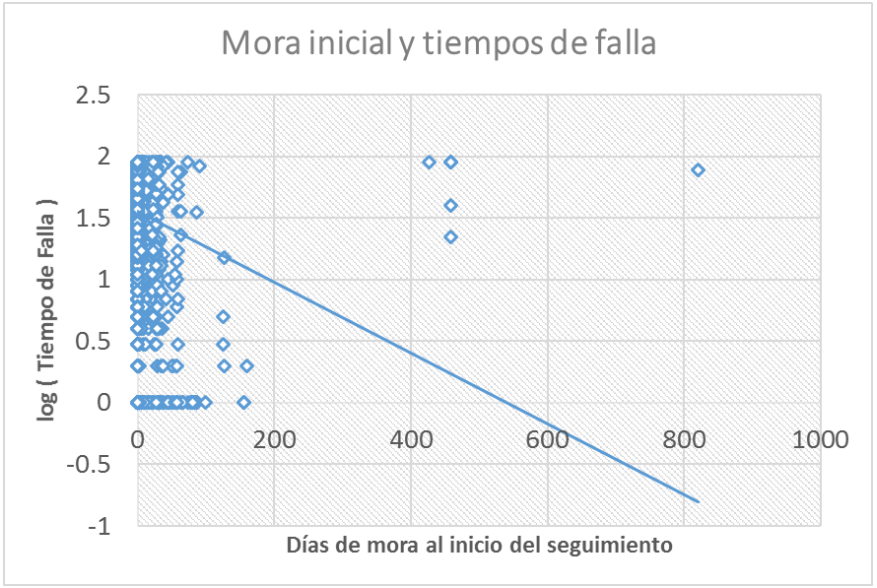
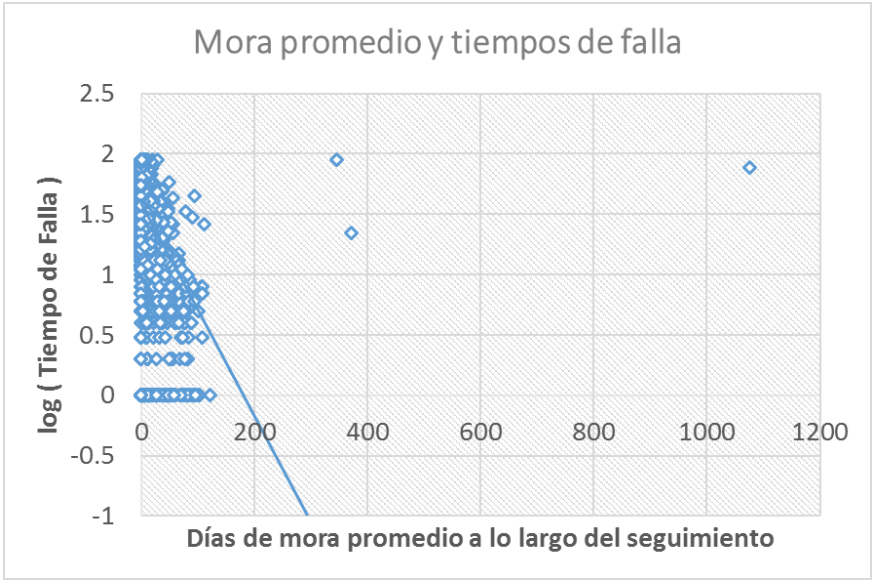


Figura 5.3: Porcentajes de incumplimiento y tiempos de falla según mora inicial y promedio.









5.2.1. Antigüedad de la cuenta corriente e incumplimiento

La Figura 5.2 (arriba) muestra los porcentajes de incumplimiento para diferentes tramos de antigüedad inicial y promedio de la cuenta corriente.

En ambos gráficos se observa que una mayor antigüedad se relaciona con un menor incumplimiento. Sin embargo, se observa un mayor decrecimiento porcentual cuando se considera la antigüedad promedio, esto es debido a que ésta última considera mayor información del cliente.

Respecto a la relación con los tiempos de falla (Figura 5.2 (abajo)), se observa que ambas antigüedades tienen un efecto positivo.

5.2.2. Mora e incumplimiento

Para los casos de la mora inicial y promedio, ilustrados en la Figura 5.3, se observa que la gran mayoría de los clientes presentan bajos niveles de mora, cantidad que va disminuyendo a medida que aumenta el retraso (Figura 5.3 (arriba)).

Por otro lado, en los primeros 3 o 4 tramos de mora hay un aumento en el porcentaje de incumplimiento, para luego disminuir ligeramente en los últimos tramos.

Este resultado es contra-intiutivo, debido a que se esperaría que aquellos clientes que superaron los 90 días de mora, sean considerados bajo incumplimiento. No obstante, el decremento del incumplimiento en el último tramo es evidencia de que existen otros criterios que también entran en juego en la declaración del *default*. A pesar de ello, la cantidad de clientes en esta situación es poca (14).

Dado que el incumplimiento aumenta en los tramos de mayor cantidad de clientes, se conjetura que la mora tiene un efecto positivo sobre el incumplimiento.

Por otro lado, al ver la dispersión entre los tiempos de vida y la mora (Figura 5.3 (abajo)), se observa una relación negativa, la cual se acentúa cuando se considera la variable en el promedio.

5.2.3. Renegociación e incumplimiento

De acuerdo con la Figura 5.4 (arriba), la mayoría de los clientes no renegocia sus créditos.

No obstante, aquellos clientes que renegociaron en algún momento de su vida (Figura 5.4 (arriba y derecha)), tienen una presencia considerable de 1331 personas, cuya cantidad representa 1/8 de los clientes que no presentaron nunca renegociación. En términos de porcentajes de incumplimiento, dentro del primer grupo, cerca de un 90 % de los clientes incumplieron; mientras que, en el segundo grupo, lo hizo cerca de un 20 %. Ello reafirma la idea concebida

en el sector bancario respecto de que la renegociación es indicativa del incumplimiento, ya que clientes más deteriorados optan por tomarlas al ver que el incumplimiento es inminente.

Para el caso de la renegociación evaluada al inicio (Figura 5.4 (arriba y izquierda)), se observa que sólo 220 clientes presentaron renegociación al inicio de su seguimiento, lo cual se atribuye a la existencia de clientes que ya habían tomado algún de consumo antes del periodo de estudio.

Por otro lado, cerca de un 30 % de los clientes que no presentaron renegociación al inicio, terminaron incumpliendo. Mientras que dicho porcentaje fue de aproximadamente un 65 % para aquellos que no lo hicieron.

Estas diferencias con respecto al caso promedio indican la desventaja que tiene utilizar la información inicial como predictor del incumplimiento, ya que la renegociación podría presentarse en algún futuro cercano, redundando en una mayor probabilidad de incumplimiento, lo cual no es captado por la información inicial. De hecho, puede haber clientes que no hayan presentado renegociación al inicio, pero pueden presentar renegociación en periodos posteriores y luego caer.

De todas maneras, estos resultados indican una relación positiva entre renegociación e incumplimiento. Asimismo, según la Figura 5.4 (abajo), se reporta una relación negativa entre la mora y el tiempo de vida.

5.2.4. Score e incumplimiento

De la Figura 5.5 (arriba), se observan diferencias notorias en la distribución de los clientes sobre los diferentes tramos de score. Si se consideran las variables al inicio, la mayoría de los clientes se concentran en el tramo de riesgo medio (2784 personas en el segmento 4). Mientras que en el caso de las variables promedio, las personas se concentran en un tramo de bajo riesgo.

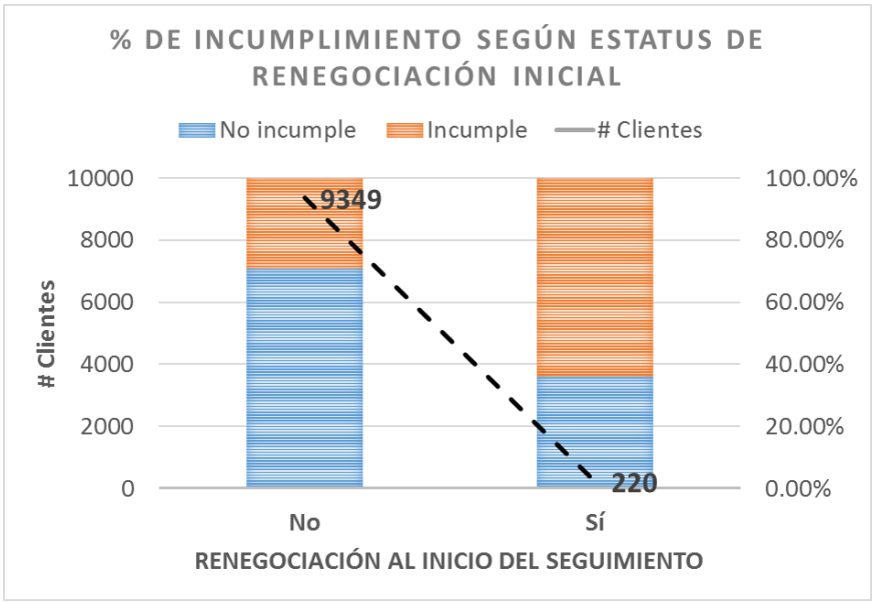
En virtud de que el score promedio resume la información de la vida total del cliente, se puede decir que la entidad tiene una cartera de clientes sana.

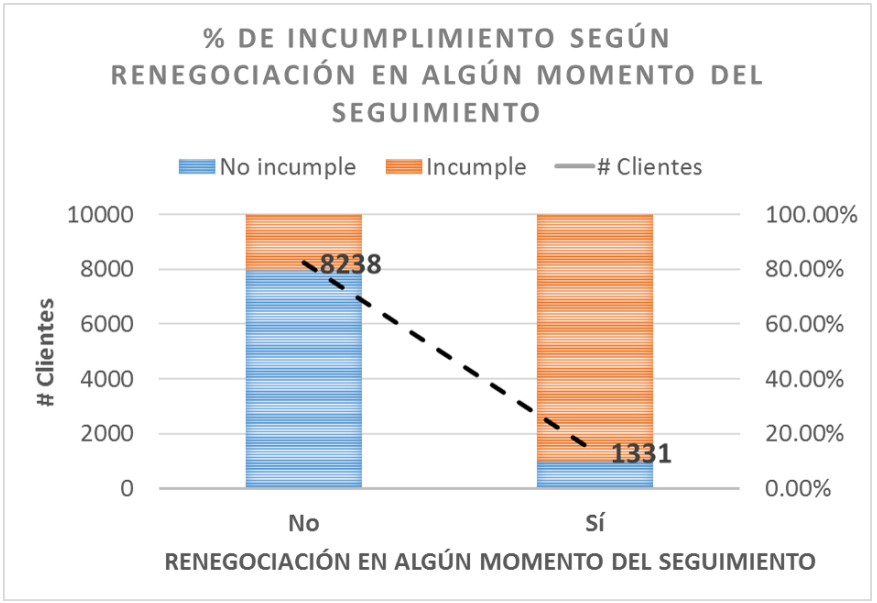
A pesar de ello, tanto score inicial y score promedio (categóricos) tienen una relación negativa con el incumplimiento. Asimismo, presentan una relación positiva con los tiempos de supervivencia 5.5 (abajo).

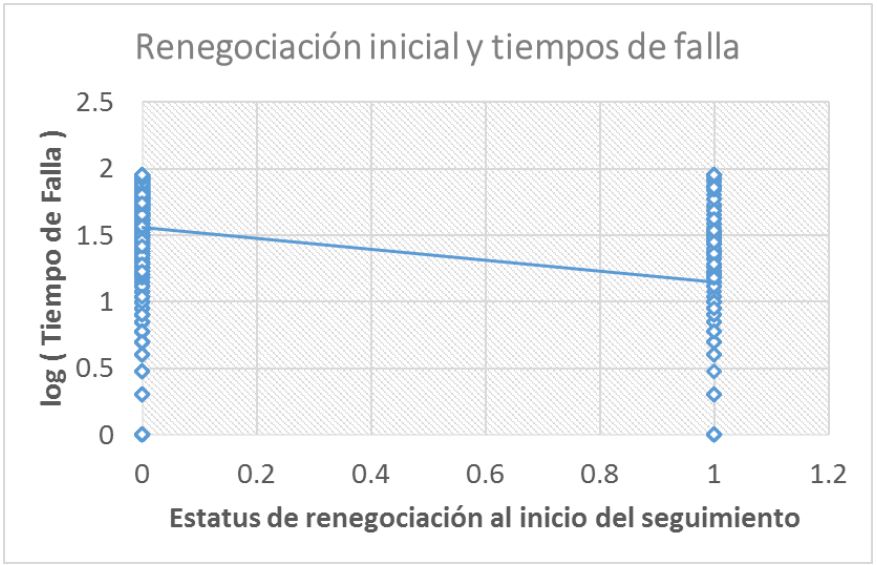
5.3. Correlaciones entre variables

Se especula que el score tiene una alta correlación con el resto de las variables, puesto que este último nace de un modelo de scoring subyacente entrenado a partir de dichas variables. En la Tabla 5.5 y en la Tabla 5.4, se reportan las correlaciones entre variables iniciales y promedio, respectivamente.

Figura 5.4: Porcentajes de incumplimiento y tiempos de falla según renegociación inicial y renegociación en algún momento del tiempo.







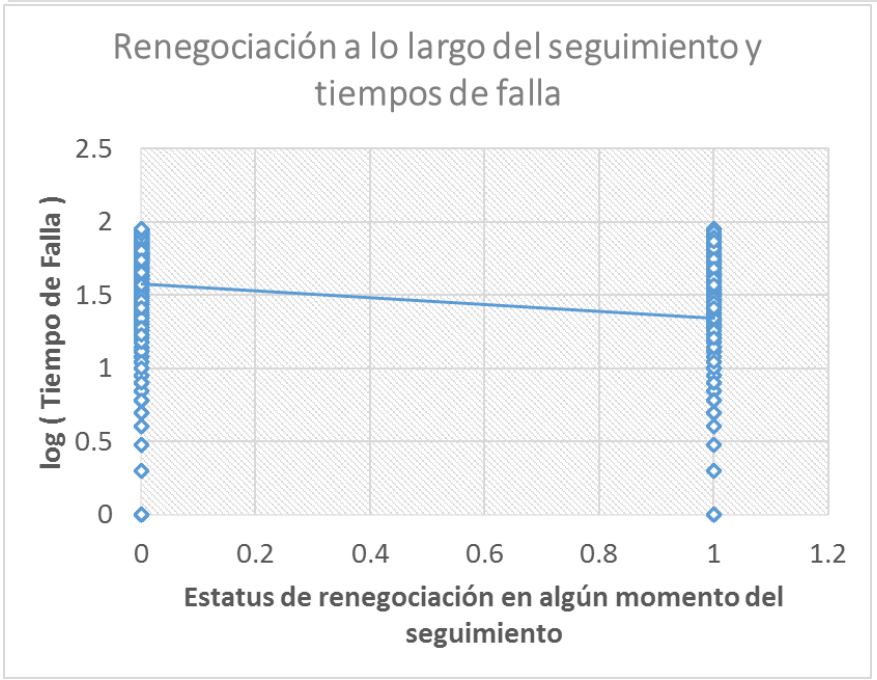
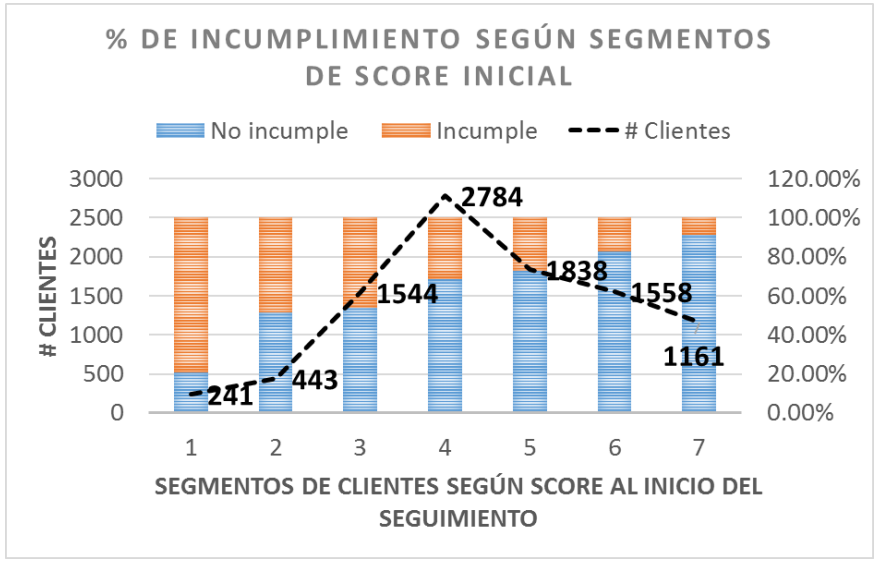
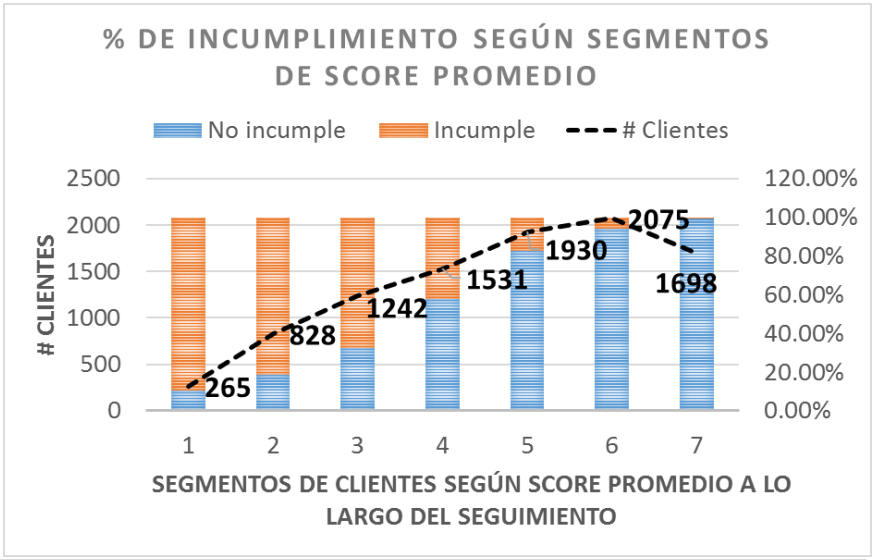
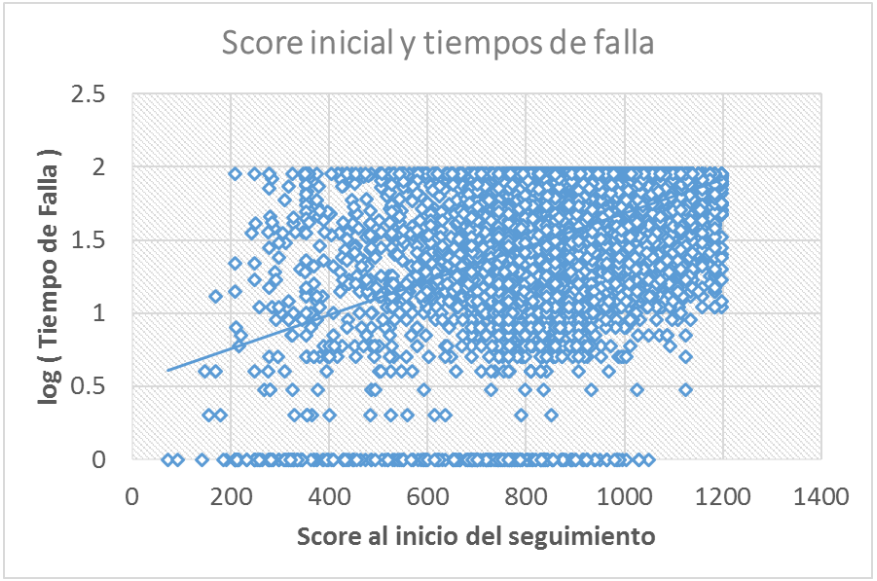


Figura 5.5: Porcentajes de incumplimiento y tiempos de falla según renegociación inicial y renegociación en algún momento del tiempo.







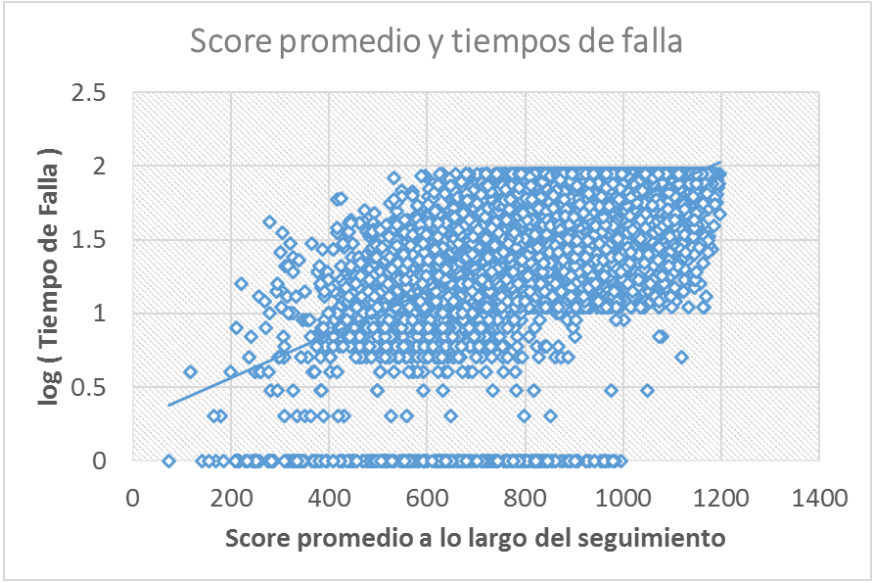


Tabla 5.4: **Matriz de correlación entre covariables promedio.** Se reporta una alta correlación del score con el resto de las covariables. Se destaca la correlación considerable entre renegociación y mora. Score se correlaciona positivamente con la antigüedad y negativamente con la mora y la renegocación.

Glosario: **antProm** = Antigüedad promedio, **moraProm** = Mora promedio, **renegDummy** = Renegociación a lo largo del seguimiento, **scoreProm** = Score promedio.

	antProm	moraProm	scoreProm	renegDummy
antProm	1.00	-0.09	0.46	-0.06
moraProm	-0.09	1.00	-0.44	0.14
scoreProm	0.46	-0.44	1.00	-0.32
renegDummy	-0.06	0.14	-0.32	1.00

Tabla 5.5: **Matriz de correlación entre covariables iniciales.** Se reporta una alta correlación del score con el resto de las covariables. Se destaca la correlación considerable entre renegociación y mora. Score se correlaciona positivamente con la antigüedad y negativamente con la mora y la renegocación.

Glosario: **antIni** = Antigüedad inicial, **moraIni** = Mora inicial, **renegIni** = Renegociación inicial, **scoreIni** = Score inicial.

	antIni	moraIni	renegIni	scoreIni
antIni	1.00	0.05	0.10	0.39
moraIni	0.05	1.00	0.12	-0.20
renegIni	0.10	0.12	1.00	-0.26
scoreIni	0.39	-0.20	-0.26	1.00

Los resultados indican que, a nivel promedio, la variable de score es la que tiene mayor correlación con el resto, con valores que superan el 30 % en valor absoluto. También se observa una correlación considerable entre la renegociación y la mora (13.88 %). Además, el score tiene relación positiva con la antigüedad y una relación negativa con la mora y la renegociación.

A nivel inicial, los resultados son análogos, salvo que la correlación del score con el resto de las covariables es ligeramente menor que en el caso promedio. Se destaca la correlación del 12 % que existe entre la mora inicial y la renegociación inicial.

Capítulo 6

Resultados

En el presente capítulo se reportan y se analizan los resultados de las regresiones AFT paramétricas y de los modelos de Cox estándar y extendido, los cuales fueron propuestos para predecir las probabilidades de incumplimiento en el tiempo.

El capítulo se divide en dos secciones: una dedicada para los modelos AFT y otra enfocada en los modelos de Cox.

En cada sección se reportan estimaciones, se grafican curvas de PI y se realizan evaluaciones estadísticas, tanto desde un enfoque gráfico como a partir de métricas. A lo largo de estas actividades, también se efectúan análisis específicos.

El objetivo de este capítulo es hacer un análisis y evaluación estadística de modelos, para dar el paso, en el Capítulo 7, a su selección de acuerdo a criterios de gestión.

6.1. Modelo AFT paramétrico

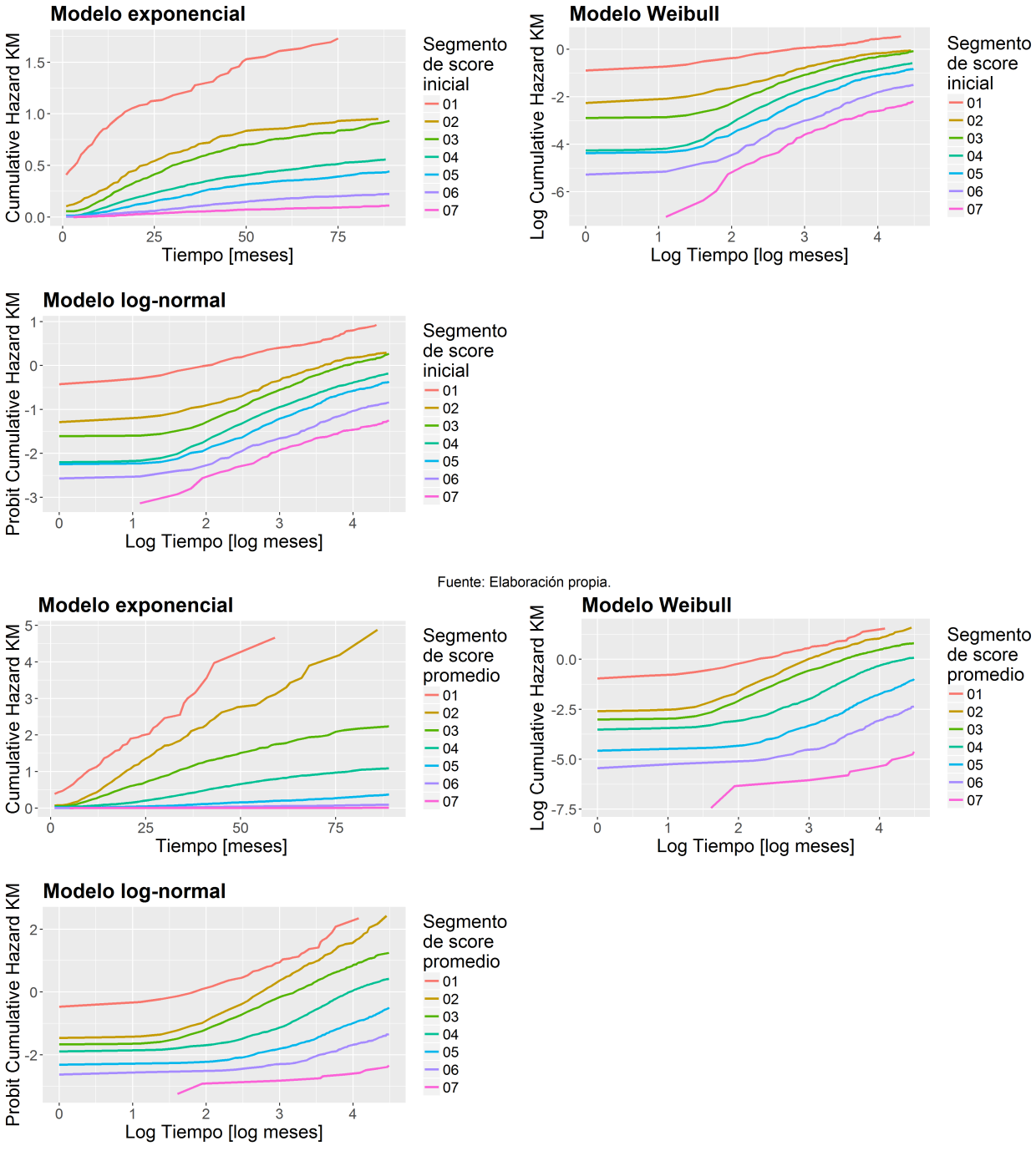
6.1.1. Gráficos de diagnóstico

En primer lugar, antes de estimar las regresiones, se mostrarán los gráficos de diagnóstico para tener una idea de qué distribución es apropiada para el tiempo de falla.

En la Figura 6.1 se reportan los gráficos de diagnóstico para los datos segmentados según score categórico inicial (los primeros 3 de arriba) y para los segmentados en base al score categórico promedio (los 3 últimos de abajo).

Los gráficos indican que para el caso inicial los mejores modelos son los Weibull y los log-normales, ya que las gráficas son lo más parecido a una recta. Mientras que para el caso promedio, el mejor modelo es el Weibull.

Figura 6.1: **Gráficos de diagnóstico para diferentes distribuciones de los tiempos de falla, según segmentos de score inicial y promedio.** Estos gráficos se logran luego de aplicar una transformación sobre la función de riesgo acumulada, de modo que ésta tenga una relación lineal con alguna función del tiempo. Posteriormente se debe debe evaluar sobre la transformación el riesgo acumulado de Kaplan-Meier (estratificado según score categórico inicial y promedio, respectivamente), con el objeto de verificar si la relación es efectivamente lineal al considerar los datos. Las distribuciones son adecuadas si las curvas son lineales. Para el caso exponencial se exige que además que la curva tenga pendiente 1.



Fuente: Elaboración propia.

Fuente: Elaboración propia.

6.1.2. Entrenamiento de los modelos

En las Tablas 6.1 e 6.2, se reportan los resultados de las regresiones AFT que consideran covariables iniciales y promedio de cada cliente, respectivamente. Cada una de las tablas presenta configuraciones que asumen distintas distribuciones, y además contemplan diferentes combinaciones de regresores.

En particular, para los modelos lognormal (5) y Weibull (8) de la Tabla 6.1, se graficaron las trayectorias de PI dentro de la Figura 6.2.

De ella cabe mencionar que el ajuste lognormal es el que más se acerca al modelo no paramétrico, que es el modelo más simple para construir las curvas de PI.

De ambos gráficos se puede observar que el segmento de mayor riesgo es el que tiene una mayor probabilidad de incumplimiento. Además, para el caso lognormal, su crecimiento es más rápido, alcanzando una PI de 90 % para el mes 90.

Para el resto de los segmentos las diferencias no son tan importantes, excepto entre los segmentos 2-3 y 4-5. Ello indica dos cosas: que hay una mayoría de clientes de buena calidad y que existen segmentos adyacentes con puntajes cercanos, sobre los cuales hay una alta concentración de clientes.

Otras formas de segmentar a la clientela sería a través de percentiles o a través de valores arbitrarios definidos por el analista. Ello podría cambiar la posición vertical de las curvas, lo cual se dejará como propuesto.

En relación a lo anterior, cabe destacar la gran influencia del score en la posición de las curvas en la Figura 6.2, lo cual se condice con los efectos reportados tanto en en la Tabla 6.1, como en la Tabla 6.2.

Finalmente, para repasar el álgebra planteada en el Capítulo 3, se construirá la PI para el segmento 2 del modelo exponencial (2) perteneciente a la Tabla 6.1:

$$\begin{aligned}
\hat{PI}_{\text{EXP}}\left(t|\overline{Z}\right) &= \hat{PI}_{\text{EXP}}\left(t\mathrm{e}^{-\overline{Z}'\hat{\beta}}|Z=0\right) \\
&= 1-\exp\left(-\hat{\lambda}\mathrm{e}^{-\overline{Z}'\hat{\beta}}t\right) \\
&= 1-\exp\left(-\hat{\lambda}\exp\left(0{,}435\cdot\overline{\text{renegIni}}-0{,}791\right)t\right)
\end{aligned}$$

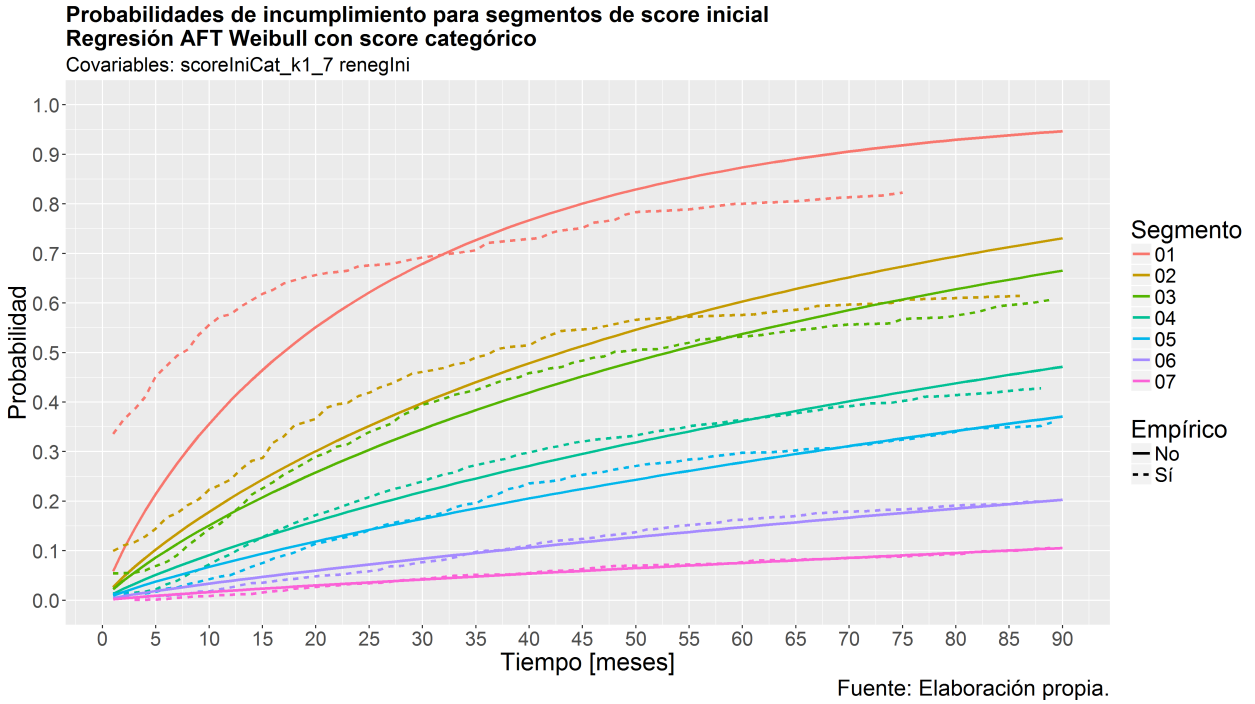
Donde:

- $\overline{Z}=\left(\overline{\text{renegIni}},1,0,0,0,0,0\right)$ corresponde al individuo promedio del segmento 2.
- $\hat{\lambda}=\mathrm{e}^{-\text{Intercepto}}=\mathrm{e}^{-3{,}416}$ es el estimador del parámetro de escala de la distribución exponencial.

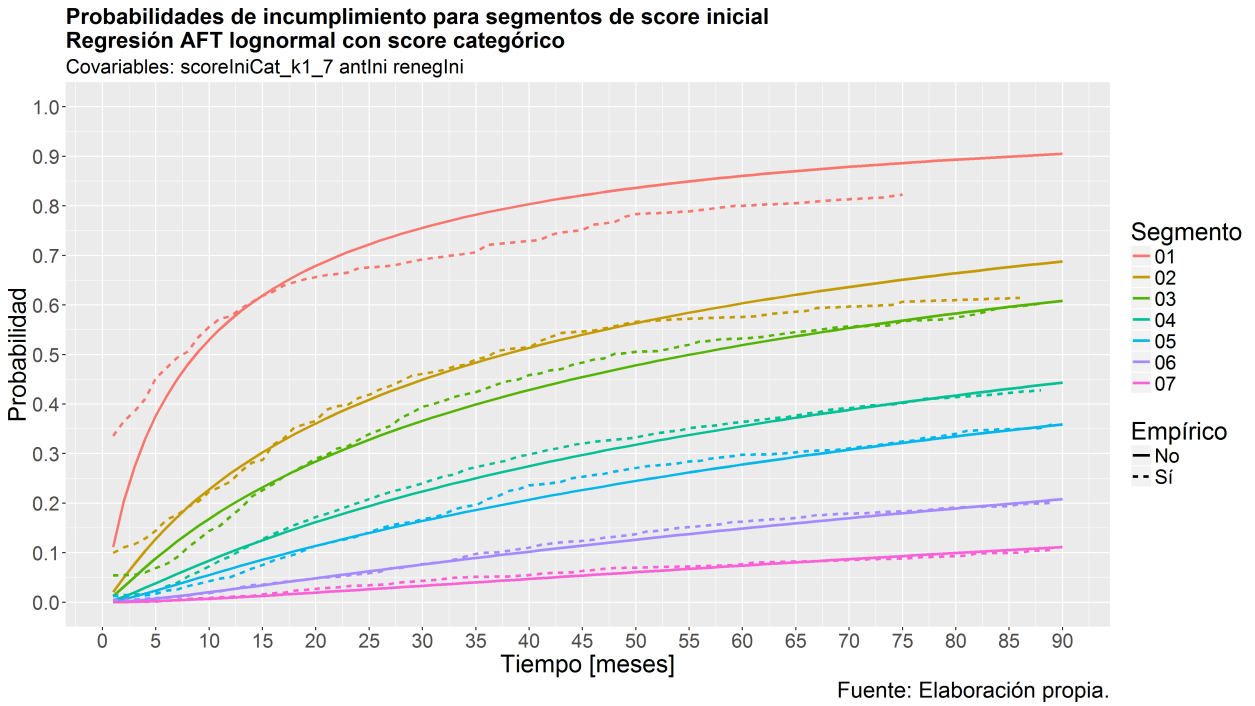
Más detalles en torno a los efectos de las covariables, se darán en el siguiente apartado.

Figura 6.2: **Probabilidades de incumplimiento crediticio modelos Weibull y lognormal con covariables iniciales.** Las curvas describen las PI paramétricas de cada segmento de clientes (línea gruesa). Dado que los modelos pronostican una PI individual, para graficar la curva de cada segmento se evaluaron las covariables en el promedio de cada grupo. Se consideraron los modelos de las columnas (5) y (8) de la Tabla 6.1. Además se incluyó en cada gráfico la PI de Kaplan-Meier a modo de referencia (línea punteada).

Glosario: **antIni** = Antigüedad inicial, **renegIni** = Renegociación inicial, **scoreIniCat_k1_7** = Score categórico inicial.



Fuente: Elaboración propia.



Fuente: Elaboración propia.

Tabla 6.1: **Resultados regresiones AFT con covariables evaluadas al inicio del seguimiento del cliente.** Se muestran resultados considerando diferentes distribuciones del tiempo de falla y distintas combinaciones de covariables. Los coeficientes positivos desaceleran el incumplimiento. Tabla elaborada con la librería *stargazer* (Hlavac, 2015).

	Variable dependiente:								
	Log tiempo de supervivencia								
	Exponencial			*Log-normal*			*Weibull*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Mora inicial	−0.001		−0.003***	−0.002		−0.011***	−0.001		−0.004***
	(0.001)		(0.0005)	(0.001)		(0.001)	(0.001)		(0.001)
Antigüedad inicial	−0.0001		0.007***	−0.003***	−0.003***	0.006***	−0.001		0.008***
	(0.0005)		(0.0004)	(0.001)	(0.001)	(0.001)	(0.001)		(0.001)
Renegociación inicial	−0.432***	−0.435***	−1.431***	−0.433***	−0.441***	−2.044***	−0.477***	−0.497***	−1.693***
	(0.092)	(0.090)	(0.088)	(0.140)	(0.140)	(0.152)	(0.107)	(0.105)	(0.110)
Segmento de score inicial 2	0.779***	0.791***		1.346***	1.378***		0.896***	0.919***	
	(0.101)	(0.100)		(0.153)	(0.151)		(0.118)	(0.116)	
Segmento de score inicial 3	0.891***	0.910***		1.579***	1.623***		1.013***	1.053***	
	(0.086)	(0.082)		(0.136)	(0.132)		(0.101)	(0.097)	
Segmento de score inicial 4	1.435***	1.453***		2.301***	2.343***		1.627***	1.667***	
	(0.085)	(0.081)		(0.134)	(0.130)		(0.101)	(0.097)	
Segmento de score inicial 5	1.782***	1.798***		2.751***	2.794***		2.010***	2.037***	
	(0.088)	(0.086)		(0.137)	(0.133)		(0.106)	(0.104)	
Segmento de score inicial 6	2.527***	2.541***		3.691***	3.738***		2.855***	2.863***	
	(0.098)	(0.096)		(0.145)	(0.140)		(0.120)	(0.117)	
Segmento de score inicial 7	3.243***	3.255***		4.517***	4.565***		3.698***	3.686***	
	(0.127)	(0.123)		(0.164)	(0.159)		(0.156)	(0.152)	
Constante	3.436***	3.416***	4.868***	2.486***	2.443***	4.936***	3.407***	3.358***	5.090***
	(0.080)	(0.074)	(0.023)	(0.128)	(0.124)	(0.040)	(0.093)	(0.086)	(0.035)
Observaciones	9,569	9,569	9,569	9,569	9,569	9,569	9,569	9,569	9,569
Log Verosim.	−16,629.260	−16,629.620	−17,229.260	−16,428.890	−16,429.650	−17,092.010	−16,585.650	−16,586.570	−17,148.550
χ^2	1,653.837***	1,653.114***	453.837***	1,685.017***	1,683.485***	358.765***	1,511.149***	1,509.309***	385.348***
	(df = 9)	(df = 7)	(df = 3)	(df = 9)	(df = 8)	(df = 3)	(df = 9)	(df = 7)	(df = 3)
Nota:									*p<0.1; **p<0.05; ***p<0.01

Tabla 6.2: **Resultados regresiones AFT con covariables evaluadas en el promedio a lo largo del seguimiento del cliente.** Se muestran resultados considerando diferentes distribuciones del tiempo de falla y distintas combinaciones de covariables. Los coeficientes positivos desaceleran el incumplimiento. Tabla elaborada con la librería *stargazer* (Hlavac, 2015).

	Variable dependiente:					
	Log tiempo de supervivencia					
	Exponencial		*Log-normal*		*Weibull*	
	(1)	(2)	(3)	(4)	(5)	(6)
Mora promedio	-0.010^{***}	-0.014^{***}	-0.011^{***}	-0.035^{***}	-0.009^{***}	-0.013^{***}
	(0.001)	(0.0005)	(0.001)	(0.001)	(0.001)	(0.0005)
Antigüedad promedio	0.001^{***}	0.014^{***}	-0.001^{***}	0.011^{***}	0.002^{***}	0.013^{***}
	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0003)	(0.0005)
Renegociación en algún momento	-0.861^{***}	-1.796^{***}	-0.747^{***}	-1.801^{***}	-0.667^{***}	-1.706^{***}
	(0.040)	(0.038)	(0.044)	(0.059)	(0.031)	(0.041)
Segmento de score promedio 2	0.581^{***}		0.804^{***}		0.516^{***}	
	(0.077)		(0.094)		(0.058)	
Segmento de score promedio 3	0.994^{***}		1.185^{***}		0.897^{***}	
	(0.077)		(0.093)		(0.058)	
Segmento de score promedio 4	1.745^{***}		2.005^{***}		1.521^{***}	
	(0.080)		(0.095)		(0.060)	
Segmento de score promedio 5	2.802^{***}		3.049^{***}		2.366^{***}	
	(0.091)		(0.100)		(0.071)	
Segmento de score promedio 6	3.962^{***}		3.929^{***}		3.194^{***}	
	(0.116)		(0.110)		(0.090)	
Segmento de score promedio 7	5.928^{***}		5.272^{***}		4.952^{***}	
	(0.268)		(0.169)		(0.237)	
Constant	2.967^{***}	4.768^{***}	2.432^{***}	4.851^{***}	2.926^{***}	4.690^{***}
	(0.075)	(0.033)	(0.091)	(0.043)	(0.055)	(0.036)
Observaciones	9,569	9,569	9,569	9,569	9,569	9,569
Log Verosimilitud	−14,160.680	−15,679.810	−14,225.260	−15,791.140	−14,051.450	−15,683.000
χ^2	$6{,}590.983^{***}$ (df = 9)	$3{,}552.724^{***}$ (df = 3)	$6{,}092.262^{***}$ (df = 9)	$2{,}960.515^{***}$ (df = 3)	$6{,}579.532^{***}$ (df = 9)	$3{,}316.445^{***}$ (df = 3)

Nota: *p<0.1; **p<0.05; ***p<0.01

Regresiones con covariables al inicio

En relación a los modelos con covariables al inicio (Tabla 6.1), se puede observar que el score, en sus distintas categorías, tiene signo positivo y es significativo al 99 % de confianza, por lo que favorece a la supervivencia. Además, sus efectos se van incrementando a medida que se avanza desde los grupos de alto a bajo riesgo (de menor a mayor dígito). Esto confirma la hipótesis inicial y va en línea con lo expresado en la exploración de datos.

Sin embargo, para los modelos (1) y (7), la antigüedad no es significativa y presenta signo negativo (favorece al incumplimiento), contrariando lo conjeturado inicialmente. No obstante, al correr los modelos (3), (6) y (9), los cuales prescinden del score categórico, se observa que la antigüedad tiene el signo esperado. Este cambio en el signo se atribuye a la alta correlación que tiene el score inicial con la antigüedad inicial (0.39), donde la primera absorbe la mayor parte del efecto de la segunda.

La mora inicial, a pesar de tener el signo esperado, en presencia del score no presenta significancia estadística. Una razón puede deberse a la correlación considerable con el score inicial (-0.20).

La renegociación es significativa al 99 %, tiene el signo esperado y, dejando de lado el score, es el regresor que tiene el mayor efecto marginal, el cual se acentúa ostensiblemente en ausencia del score.

Para los modelos AFT lognormales (4) y (6), uno con presencia de score categórico y otro sin él, se graficaron las variaciones de la PI ante cambios en cada covariable, lo cual se reporta en la Figura 6.3. De ella se constatan resultados concordantes con la Tabla 6.1, aunque la renegociación tiene un efecto limitado debido a su rango tan acotado (0 y 1).

Por otro lado, se puede contemplar cómo la presencia del score inhibe el efecto de las demás covariables, los cuales se acentúan en su ausencia. El cambio más notorio es el experimentado por la mora. Además, de ambas figuras, se visualiza el cambio de signo de la antigüedad.

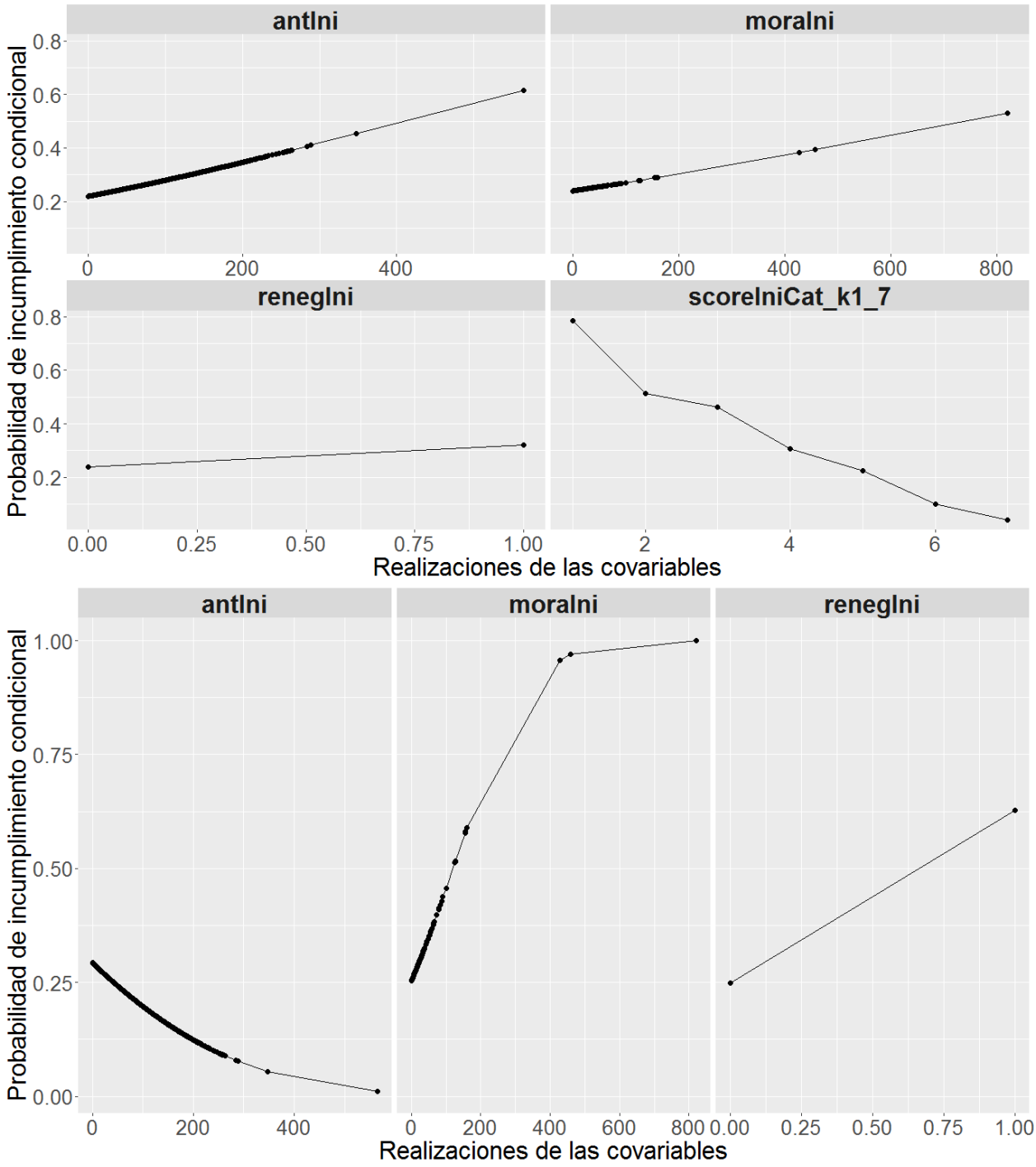
Dados los problemas que reviste juntar el score con el resto de las covariables, se recomienda estimar el modelo sólo con el score, o bien, sólo con el resto. De este modo se evitarían los problemas de correlación antes descritos.

Otro problema puede suscitarse al incorporar el score en el modelo: una posible causalidad reversa. Esto se debe a que el score nace como la linealización de una probabilidad de incumplimiento a 12 meses, la cual fue predicha a través de un modelo logístico. Luego, si se desea estimar una nueva PI (pero temporal) a través de este score, se estaría haciendo el proceso inverso.

Una forma de solucionar este problema es estimar múltiples modelos diferentes para cada grupo de score, los cuales pueden -o no- incorporar otras covariables. De todos modos, se seguirá adelante con el análisis utilizando el score categórico como covariable, pero con la advertencia ya puesta en el tapete.

Figura 6.3: **Efectos de las covariables sobre las probabilidades de incumplimiento.** Se consideró el resultado de dos modelos de distribución log-normal: uno con presencia de score categórico y otro sin él. Se describen los efectos de cada covariable (ceteris paribus) sobre la probabilidad de incumplimiento en un momento puntual del tiempo. Al variar una covariable, las demás se sitúan en el promedio. El tiempo elegido corresponde a la mediana de la muestra (43 meses).

Glosario: **antIni** = Antigüedad inicial, **moraIni** = Mora inicial, **renegIni** = Renegociación inicial, **scoreIniCat_k1_7** = Score categórico inicial.



Regresiones con covariables en el promedio

Ante el hecho de que los modelos con covariables al inicio no capturan la historia del cliente tras su ingreso, se exploró la estimación de modelos con covariables evaluadas en el promedio (ver Tabla 6.2). Es decir, para cada cliente se tomó el promedio de cada variable a lo largo de su vida, y con toda esa información se estimaron modelos AFT.

Los resultados indican que todos los regresores son significativos y poseen los signos esperados. No obstante, sus efectos económicos son más marcados respecto de los modelos con covariables iniciales, lo cual se atribuye al hecho de que las covariables promedio incorporan información de la historia de los clientes.

Sin embargo, cabe indicar que este tipo de modelos pueden carecer de sentido, por el hecho de que se están pronosticando probabilidades con información futura.

Si el interés consiste en tratar de incorporar la evolución de los clientes en los modelos, existe la posibilidad de entrenar modelos con covariables dependientes del tiempo: si bien la teoría permite estimar modelos paramétricos con covariables dependientes del tiempo, no hay paquetes estadísticos en R que permitan esto hasta el momento. Sin embargo, sí existen paquetes para el caso de los modelos semi-paramétricos, lo cual se abordará en la sección 6.2.

6.1.3. Evaluación de modelos

En la Tabla 6.3 se consignan las métricas AIC de los mejores modelos AFT paramétricos. En ella se observa que los modelos con covariables promedio tienen una mayor verosimilitud ajustada por cantidad de covariables. Sin embargo, los residuos de Cox-Snell (Figura 6.4) indican lo contrario, puesto que los ajustes con covariables promedio presentan mayores desviaciones a la identidad que los modelos evaluados con covariables al inicio.

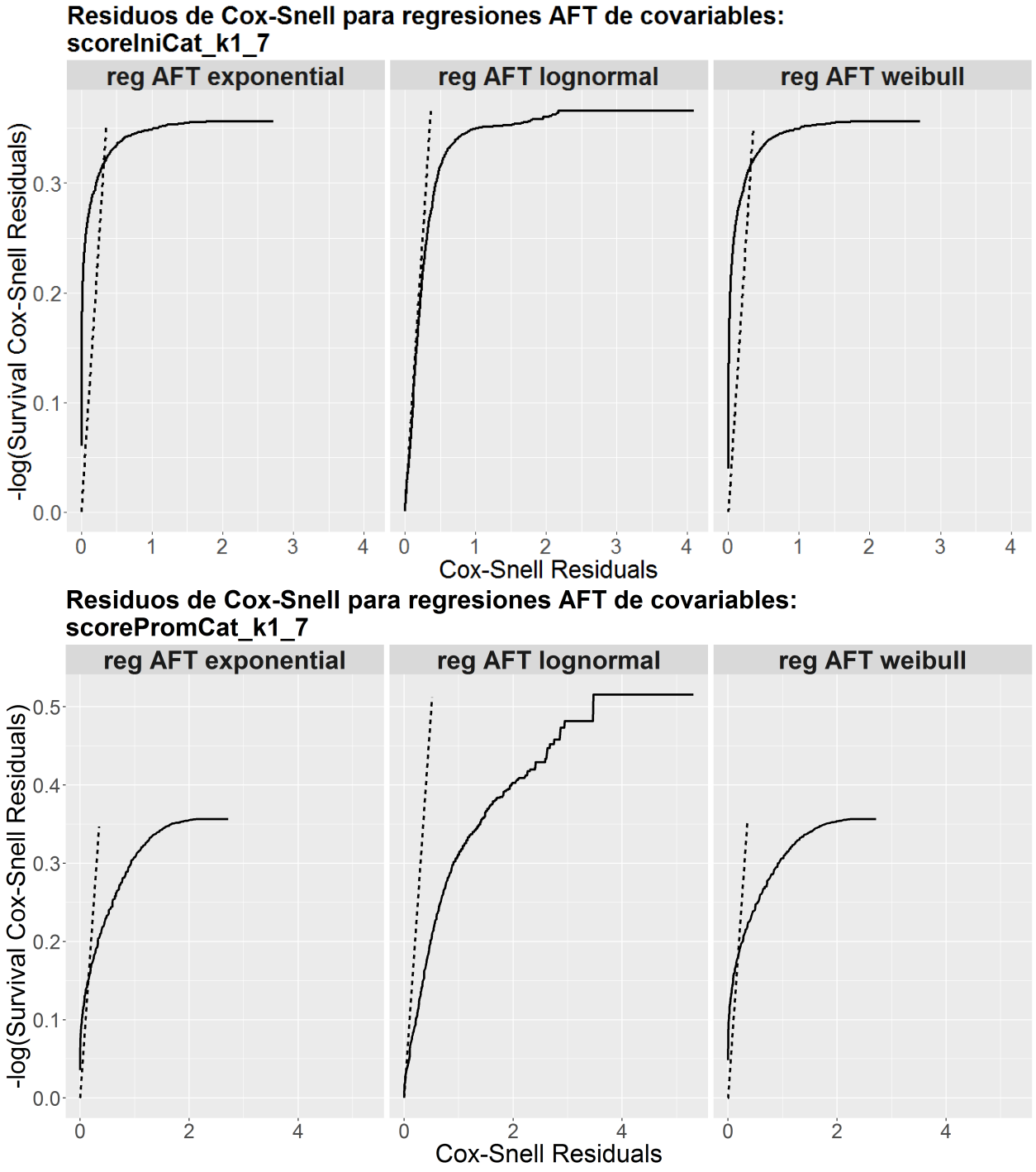
Tabla 6.3: **AIC para modelos AFT paramétricos.** Se compara los mejores modelos de cada clase (i.e. covariables promedio y al inicio).

Glosario: **antIni** = Antigüedad inicial, **renegIni** = Renegociación inicial, **scoreIniCat** = Score categórico inicial, **antProm** = Antigüedad promedio, **renegDummy** = Renegociación en algún momento del seguimiento, **scorePromCat** = Score categórico promedio, **moraProm** = Días de mora promedio

	Modelo	df	AIC	Covariables
1	Regresión AFT WEI	11	28125	`scorePromCat, moraProm, antProm, renegDummy`
2	Regresión AFT EXP	10	28341	
3	Regresión AFT LOGN	11	28472	
4	Regresión AFT LOGN	10	32879	`scoreIniCat, antIni, renegIni`
5	Regresión AFT WEI	9	33191	`scoreIniCat, antIni, renegIni`
6	Regresión AFT EXP	8	33275	

Figura 6.4: **Residuos de Cox-Snell para modelos de múltiples distribuciones con covariables iniciales y promedio.** Los gráficos describen la función de riesgo acumulada de Cox-Snell y la compara con el riesgo acumulado de una distribución exponencial de tasa 1 (la identidad). Si los puntos se acercan a la identidad, significa que el modelo es bueno. En este caso se optó por las formulaciones que sólo poseen score categórico como covariables (los resultados son similares si se ocupa otra combinación de covariables). Para la demostración de la propiedad de los residuos de Cox-Snell revisar Collett (2015).

Glosario: **scoreIniCat_k1_7** = Score categórico inicial, **scorePromCat_k1_7** = Score categórico promedio.



Cabe señalar que el AIC usualmente se ocupa como un indicador de comparación relativa, es decir, señala qué modelo es mejor qué otro, pero generalmente no es usado como un indicador absoluto de bondad de ajuste. Por otro lado, los residuos de Cox-Snell, sí son una medida absoluta de bondad de ajuste, por lo que el autor se inclina más a los modelos con covariables iniciales.

No obstante, al evaluar desde un punto de vista operativo, se pueden distinguir distintos usos de los modelos. En particular, las regresiones con covariables al inicio son adecuadas para modelos de originación, ya que la evaluación de la PI de un nuevo cliente se realizaría en base a un modelo entrenado con individuos que compartieron su condición. Sin embargo, este método no es adecuado para modelos de seguimiento, puesto que omite información del cliente.

En relación a los modelos con covariables promedio, éstos no son adecuados para modelos de originación, dado que el nuevo cliente se estaría comparando con clientes con cierta historia en el banco. Además, por el hecho de predecir PIs con información futura, no son adecuados para el seguimiento.

Finalmente, es preciso indicar que los modelos AFT entregan resultados a nivel de individuo, por lo tanto, para generar segmentos de score, irremediablemente se deben fijar covariables en el promedio del grupo, lo cual puede ser una desventaja en caso de haber otras variables dummies aparte del score (se generarían valores sin sentido). Por lo tanto, una alternativa a ello sería definir tantos modelos como segmentos de score haya, lo cual quedará propuesto.

Tabla 6.4: **Resultados regresiones de Cox con covariables evaluadas al inicio.** Los modelos están divididos en 3 grupos: aquellos que presentan el score categórico como covarible, aquellos que lo ocupan como variable de estratificación y el modelo que carece de score. El segundo grupo de modelos toman el nombre de estratificados, puesto que se definen diferentes *baseline hazard* para las submuestras inducidas por el score categórico. En cambio, los modelos de los grupos 1 y 3, presentan una *baseline* común para toda la muestra. Se muestran resultados considerando diferentes combinaciones de covariables. Tabla elaborada con la librería *stargazer* (Hlavac, 2015)

	Modelos de Cox con covariables al inicio				
	Score categótico como regresor		Score categótico como variable de estratificación		(Sin score)
	(1)	(2)	(3)	(4)	(5)
Mora inicial	0.001		0.0004		0.003***
	(0.001)		(0.001)		(0.0005)
Antigüedad inicial	0.001**	0.001**	0.001***	0.001***	−0.005***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0004)
Renegociación inicial	0.402***	0.401***	0.303***	0.303***	1.352***
	(0.092)	(0.092)	(0.093)	(0.093)	(0.088)
Segmento de score inicial 2	−0.753***	−0.763***			
	(0.101)	(0.100)			
Segmento de score inicial 3	−0.859***	−0.873***			
	(0.086)	(0.085)			
Segmento de score inicial 4	−1.385***	−1.398***			
	(0.086)	(0.084)			
Segmento de score inicial 5	−1.691***	−1.705***			
	(0.089)	(0.087)			
Segmento de score inicial 6	−2.393***	−2.409***			
	(0.098)	(0.097)			
Segmento de score inicial 7	−3.131***	−3.148***			
	(0.127)	(0.126)			
Observaciones	9,569	9,569	9,569	9,569	9,569
Cox-Snell R^2	0.136	0.136	0.002	0.002	0.034
Max. C-S R^2	0.995	0.995	0.984	0.984	0.995
Log Verosim.	−24,598.400	−24,598.710	−19,778.690	−19,778.800	−25,130.890
Wald Test	1,324.950***	1,324.180***	22.300***	22.050***	370.030***
	(df = 9)	(df = 8)	(df = 3)	(df = 2)	(df = 3)
LR Test	1,397.823***	1,397.201***	20.794***	20.572***	332.835***
	(df = 9)	(df = 8)	(df = 3)	(df = 2)	(df = 3)
Score (Logrank) Test	1,696.935***	1,695.866***	22.430***	22.176***	402.649***
	(df = 9)	(df = 8)	(df = 3)	(df = 2)	(df = 3)

Nota: *p<0.1; **p<0.05; ***p<0.01

6.2. Regresiones de Cox

6.2.1. Entrenamiento de los modelos de Cox estándar

En la Tabla 6.4 se reporta la estimación de múltiples modelos de Cox estándar. En ella existen tres grupos de modelos: aquellos que ocupan el score categórico como regresor, aquellos que lo ocupan como variable de estratificación y un modelo que carece de score en su formulación. Se entiende por estratificación al hecho de asumir distintas funciones de riesgo basal (i.e. $h_{0,j}(t)$) para diferentes segmentos de score.

En particular, para los modelos (2) y (4) de la Tabla 6.4 se graficaron las trayectorias de PI dentro de la Figura 6.5.

Para esbozar parte del cálculo, considere el modelo (2) y un cliente del segmento 2 con sus covariables evaluadas en el promedio (dicho cliente será el representante del segmento). Con dicha información, la PI estará dada por:

$$\hat{PI}_{\text{COX}}\left(t|\overline{Z}\right) = 1 - \hat{S}_0\left(t\right)^{\exp\left(\overline{Z}'\hat{\beta}\right)} \tag{6.1}$$

Donde:

- $\overline{Z} = \left(\overline{\text{antIni}}, \overline{\text{renegIni}}, 1, 0, 0, 0, 0, 0\right)$ corresponde al individuo promedio del segmento 2.
- $\overline{Z}'\hat{\beta} = 0{,}001 \cdot \overline{\text{antIni}} + 0{,}401 \cdot \overline{\text{renegIni}} - 0{,}763$
- $\hat{S}_0 = -\ln \hat{H}_0(t)$ la función de supervivencia estimada, con $\hat{H}_0$ el estimador de la función de riesgo acumulada de Breslow expresada en el Capítulo 3.

Para el caso del modelo (4), que es estratificado, se puede hacer el mismo ejercicio:

$$\hat{PI}_{\text{COX}}\left(t|\overline{Z}\right) = 1 - \hat{S}_{0,2}\left(t\right)^{\exp\left(\overline{X}'\hat{\beta}\right)} \tag{6.2}$$

Donde:

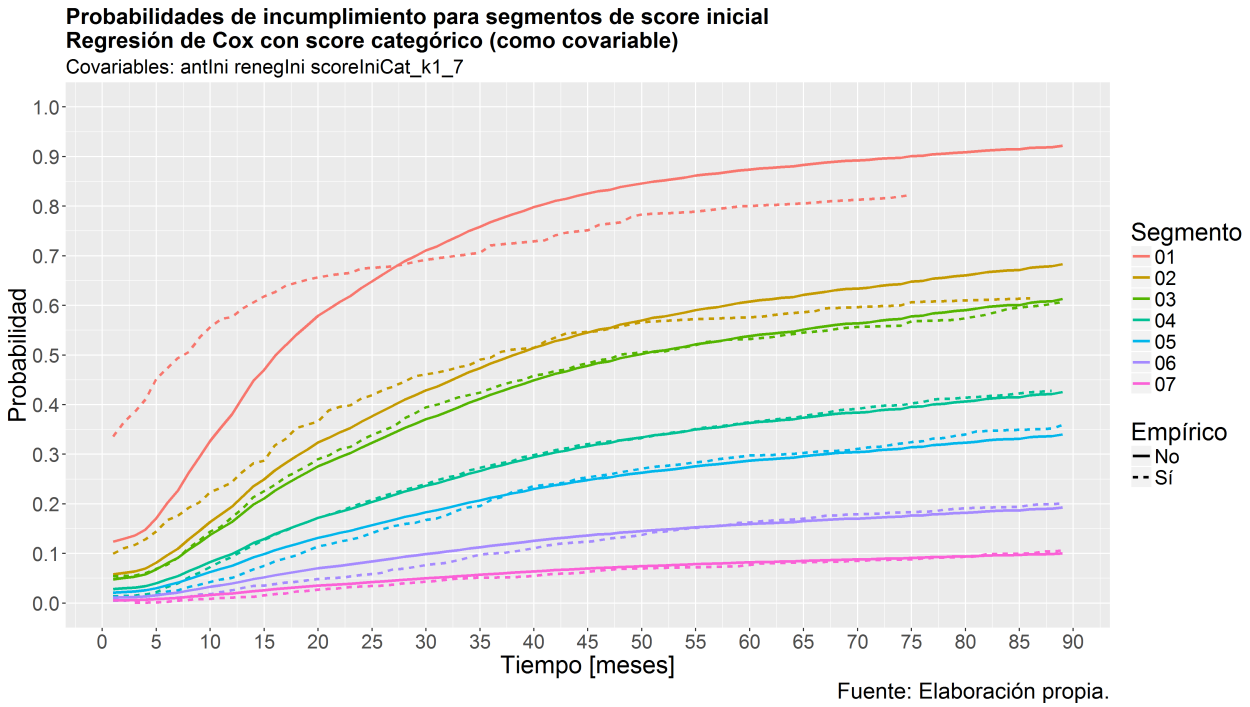
- $\hat{S}_{0,j}$ es la función de supervivencia basal del segmento j.
- $\overline{X} = \left(\overline{\text{antIni}}, \overline{\text{renegIni}}\right)$ son las covariables que hacen referencia al cliente promedio del segmento 2.

En este caso la pertenencia al segmento afecta al término base ($\hat{S}_{0,2}$) y no a la función de riesgo relativo (i.e. $\exp(Z'\beta)$). Ello se realiza calculando múltiples estimadores de Breslow para cada submuestra inducida por el score categórico.

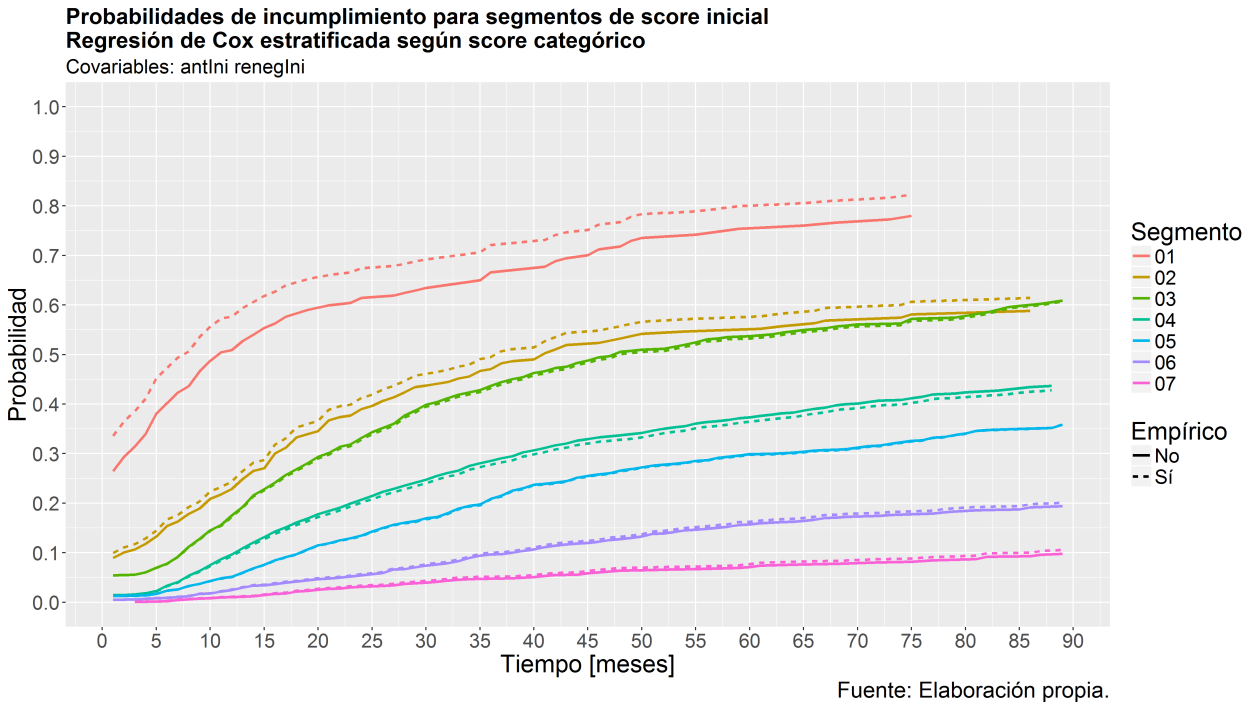
Ambas predicciones se hacen automáticamente en R, aplicando el comando `survfit` sobre el objeto `coxph` que alberga la información de las regresiones de Cox.

Figura 6.5: **Probabilidades de incumplimiento crediticio para modelos de Cox estándar.** Se describen las curvas de PI para modelos de Cox evaluados con covariables iniciales, para los cuales se incluyó el score categórico como regresor (modelo (2) en Tabla 6.4) y como variable de estratificación (modelo (4) en Tabla 6.4). Se entiende por estratificación al hecho de estimar diferentes PIs *baseline* para cada submuestra inducida por el score categórico. Las trayectorias de los modelos se diagraman con línea gruesa, mientras que las de Kaplan-Meier se hacen con línea punteada.

Glosario: **antIni** = Antigüedad inicial, **renegIni** = Renegociación inicial, **scoreIniCat_k1_7** = Score categórico inicial.



Fuente: Elaboración propia.



Fuente: Elaboración propia.

De la Figura 6.5 se observa que ambos gráficos no tienen una forma suave, como lo es en el caso de los modelos paramétricos. Esto se debe a que existe una parte del modelo que se estima de forma no paramétrica.

Sin embargo, es preciso indicar las diferencias entre ambos modelos de Cox. En particular, el gráfico del modelo estratificado (abajo) está prácticamente sincronizado con el estimador de Kaplan-Meier. Aunque dependiendo del segmento de riesgo, está sistemáticamente sobre o bajo dicha curva.

La sincronía se explica porque tanto el estimador de Kaplan-Meier como el modelo de Cox, están estratificados según el mismo score categórico, es decir, la función de supervivencia basal estimada (i.e. $\hat{S}_0(t)$) es prácticamente la misma para ambos segmentos a lo largo del tiempo (las únicas diferencias podrían derivarse del hecho de que la baseline del modelo se Cox se obtuvo mediante la aproximación de Breslow).

Mientras que las distancias se deben al efecto del riesgo relativo del cliente evaluado (i.e. el valor $\exp(Z'\hat{\beta})$ asociado al cliente promedio de cada segmento). De hecho, siempre se espera que exista una distancia entre ambos modelos; si ésta es positiva o negativa dependerá de la información del cliente y de los parámetros estimados.

Para el gráfico con el score categórico como regresor (arriba), no existe la sincronía debido a dos razones: el modelo semi-paramétrico posee una única *baseline hazard* para toda la muestra (i.e. $h_0(t)$) y, en segundo lugar, método no paramétrico posee múltiples *hazards* (i.e. $h_{0,j}(t)$ con $j \in 1, \ldots, J$). Ello indicaría las diferencias considerables entre los pronósticos no paramétricos y semi-paramétricos en los primeros meses de seguimiento.

Otro aspecto a considerar es el cruzamiento entre el segmento 2 y el 3, del cual se descarta que sea producto de las características de los clientes, puesto que las estimaciones no-paramétricas también se cruzan. Una recomendación sería que ambos segmentos se fusionaran.

Por otro lado, a lo largo de los segmentos 2 al 7, no existen diferencias sustanciales de PI entre los dos modelos y el método no paramétrico. Sin embargo, sí se reporta una diferencia importante en el segmento 1 (de mayor riesgo), lo que levanta las sospechas de que el modelo con score como covariable podría tener menor calidad que el resto. Ello se abordará debidamente en la sección de evaluación de modelos.

Efectos de las covariables

Volviendo a la Tabla 6.4, se observa que todos los signos de las covariables son los esperados, excepto en el caso de la antigüedad cuando se presenta el score como regresor o como variable de estratitifación (modelos (1) y (4))[1].

Todas las covariables son significativas al menos en un 95 % de confianza, excluyendo a la

[1]Recordar que en los modelos de Cox, a diferencia de los modelos AFT paramétricos, un signo positivo indica que la covariable favorece al incumplimiento.

mora, cuando el score actúa como regresor o variable de estratificación.

El crecimiento de las magnitudes de score a lo largo de las categorías, es concordante con el riesgo asociado a cada segmento. Además, dicha variable es la que genera los mayores efectos marginales. En un segundo puesto entra la renegociación.

En ausencia del score (modelo (5)), el resto de las covariables aumentan considerablemente sus efectos marginales, presentan los signos correctos y son significativas al 99 % de confianza. Sin embargo, en términos del pseudo R2 de Cox-Snell, el modelo presenta una baja bondad de ajuste, lo cual puede deberse a que existen variables que se están omitiendo que pueden ser de relevancia en el pronóstico.

Además, a pesar de que la mora sea significativa en este modelo, su efecto económico es bajo. Lo cual iría en contra-sentido con la definición de incumplimiento (i.e. 90 días de mora en los próximos 12 meses, entre otros criterios más).

Una posible causa deriva del hecho de que clientes con mora inicial no nula, deben ser clientes que han estado vigentes antes de la fecha de estudio. De lo contrario se esperaría que presentaran mora inicial de 0 días, a menos que el monitoreo sea imperfecto.

Los clientes que cumplen con esta condición son denominados *datos truncados* (Kalbfleisch y Prentice, 2002; Klein y Moeschberger, 2005), los cuales pueden generar sesgo en la estimación.

En presencia de truncamiento existen versiones de la función de verosimilitud general (recordar la expresión 3.2.2 en el Capítulo 3) que permiten dar cuenta del fenómeno. Por lo que aplicar ello sería una primera recomendación.

Una segunda recomendación sería borrar de la muestra a todos los clientes con vigencia al inicio del estudio, de este modo se evita el truncamiento, pero con el costo de perder variabilidad.

Tomando en cuenta lo anterior, y añadiendo el cumplimiento del supuesto de clientes iniciales con mora 0, entonces cabría preguntarse si la mora, de por sí, es un buen indicador para pronosticar la PI. Esta pregunta quedará pendiente para la sección de modelos de Cox extendidos.

Finalmente, de acuerdo a los tres test de significancia global, todos los modelos son estadísticamente significativos al 99 % de confianza.

Test de riesgos proporcionales

Recordar que el modelo de Cox estándar descansa sobre el supuesto de que la *hazard rate* entre individuos cualquiera dentro de la muestra, es constante.

Esto equivale a plantear:

$$HR = h_0(t)\exp(Z_1'\beta)/h_0(t)\exp(Z_2'\beta) = \exp((Z_1 - Z_2)'\beta) = \text{cte} \tag{6.3}$$

Donde Z_1 y Z_2 son las características de dos clientes arbitrarios dentro de la muestra y β el conjunto de covariables del modelo, las cuales se asumen como únicas para todos los individuos.

Cabe señalar que este ejercicio se puede hacer por variable variable, como también para todo el modelo en su conjunto.

Para ello se recurrirá a al test de Grambsch y Therneau (1994) y al análisis de los residuos de Schoenfeld en el tiempo, los cuales se aplicarán sobre el modelo (1) de la Tabla 6.4.

Tabla 6.5: **Test de Grambsch y Therneau para riesgos proporcionales.** Se testea la hipótesis nula de funciones de riesgo proporcionales a través de un estadístico χ^2 de un grado de libertad para cada covariable. Para el test global se utiliza un estadístico χ^2 de p grados de libertad, donde p es el número de covariables.

Covariable	χ^2	p-valor
Mora inicial	7.274	0.007
Antigüedad inicial	6.16	0.013
Renegociación inicial	0.882	0.348
Score categórico inicial 2	10.168	0.001
Score categórico inicial 3	37.421	1.E-09
Score categórico inicial 4	54.038	2.E-13
Score categórico inicial 5	81.033	0.E+00
Score categórico inicial 6	92.782	0.E+00
Score categórico inicial 7	60.463	8.E-15
GLOBAL	189.866	0.00E+00

De la Tabla 6.5 se deduce que la renegociación es la única covariable para la cual no se rechaza el supuesto (p-valor $>$ 0,04). Aunque para tener seguridad, se verán los gráficos de los $\beta(t)$ basados en los residuos de Schoenfeld, los cuales se muestran en las Figuras 6.6, 6.7 y 6.6.

Los resultados indican que las variables de score son las que incumplen con el supuesto, puesto que existen regiones donde la curva se aleja de una pendiente 0. Lo mismo se puede decir de la antigüedad.

Para remediar ello existen dos opciones: estratificar las covariables conflictivas o emplear un modelo que por definición admita la violación de los riesgos proporcionales.

La primera opción ya se realizó al entrenar los modelos estratificados (3) y (4) de la Tabla 6.4. Por lo que a continuación se evaluará la segunda opción por medio de modelos de Cox extendidos.

Entonces la aplicación de modelos de Cox con covariables dependientes del tiempo no sólo permite ocupar la historia del cliente en la estimación, sino que también se hace cargo del incumplimiento del supuesto de riesgos proporcionales.

Tabla 6.6: **Resultados regresiones de Cox con covariables dependientes del tiempo.** Los modelos están divididos en 3 grupos: aquellos que presentan el score categórico como covariable, aquél que no ocupa score, y aquellos que ocupan el score categórico como variable de estratificación. Tabla elaborada con la librería *stargazer* (Hlavac, 2015).

	Modelos de Cox con covariables dependientes del tiempo				
	Score categótico como regresor		(Sin score)	Score categótico como variable de estratificación	
	(1)	(2)	(3)	(4)	(5)
Mora d.d.t.	0.001***	0.001***	0.003***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Antigüedad d.d.t.	−0.001		−0.006***	−0.001*	
	(0.0004)		(0.0004)	(0.0004)	
Antigüedad inicial		−0.002***			−0.001***
		(0.0005)			(0.0004)
Renegociación d.d.t.	1.015***	1.043***	2.225***	1.191***	1.213***
	(0.069)	(0.069)	(0.067)	(0.070)	(0.070)
Segmento de score d.d.t. 2	−1.700***	−1.703***			
	(0.059)	(0.059)			
Segmento de score d.d.t. 3	−2.666***	−2.674***			
	(0.075)	(0.075)			
Segmento de score d.d.t. 4	−3.618***	−3.625***			
	(0.076)	(0.076)			
Segmento de score d.d.t. 5	−4.652***	−4.655***			
	(0.101)	(0.101)			
Segmento de score d.d.t. 6	−5.890***	−5.881***			
	(0.152)	(0.152)			
Segmento de score d.d.t. 7	−7.482***	−7.455***			
	(0.291)	(0.291)			
Observaciones	234,364	234,364	234,364	234,364	234,364
Cox-Snell R^2	0.050	0.050	0.006	0.001	0.001
Max. C-S R^2	0.194	0.194	0.194	0.131	0.131
Log Verosim.	−19,268.570	−19,263.800	−24,595.880	−16,251.150	−16,247.420
Wald Test	7,363.250***	7,377.080***	3,018.940***	375.370***	382.690***
	(df = 9)	(df = 9)	(df = 3)	(df = 3)	(df = 3)
LR Test	12,031.290***	12,040.830***	1,402.851***	279.434***	286.904***
	(df = 9)	(df = 9)	(df = 3)	(df = 3)	(df = 3)
Score (Logrank) Test	37,930.190***	37,935.690***	13,459.900***	420.638***	428.069***
	(df = 9)	(df = 9)	(df = 3)	(df = 3)	(df = 3)

Nota: *p<0.1; **p<0.05; ***p<0.01

6.2.2. Entrenamiento de los modelos de Cox extendidos

Si bien tanto la mora como la renegociación cumplen con el supuesto de riesgo proporcionales (en base al criterio gráfico), de todas maneras de utilizarán como covariables dependientes del tiempo. Lo mismo aplicará para la antigüedad y para el score categórico.

La Tabla 6.6 muestra el resultado de entrenar múltiples modelos de Cox extendidos.

De ella se puede observa que se agregó también la antigüedad inicial, lo cual responde a dos razones: en primer lugar, porque en este tipo de modelo ya no es de interés preservar el supuesto de riesgos proporcionales; y, en segundo lugar, porque como la antigüedad en el tiempo es totalmente predecible, el modelo -por construcción- no capturará los efectos por cada nuevo año que se añade. Por lo tanto, tendrá una misma verosimilitud parcial.

Para ilustrar este punto, se citará una adaptación al ejemplo expuesto por T. Therneau y cols. (2016). En él se considera un cliente i que incumplirá en el tiempo t_i y otro cliente j que no lo hará, por lo tanto, formará parte del conjunto de riesgo en t_i, es decir, aquellos que no han incumplido justo antes de t_i (notar que el cliente i también pertenece a ese conjunto). Sin pérdida de generalidad, para ambos se considerará como único regresor la antigüedad.

Como la antigüedad es totalmente predecible, se tendrá que: $\text{ant} = \text{antIni} + t_\text{i}$. Es decir, la antigüedad actual es la suma de la antigüedad inicial más el tiempo transcurrido hasta el incumplimiento del cliente i.

Ahora si se expresa la verosimilitud parcial marginal del cliente i se tendrá que los efectos del tiempo se cancelarían, puesto que todos los clientes experimentan el mismo efecto.

$$\frac{\exp(\beta(\text{antIni}_\text{i} + t_\text{i}))}{\sum_{j \in R(t_\text{i})} \exp(\beta(\text{antIni}_j + t_\text{i}))} = \frac{\exp(\beta \cdot \text{antIni}_\text{i})}{\sum_{j \in R(t_\text{i})} \exp(\beta \cdot \text{antIni}_j)} \tag{6.4}$$

Como el cliente i es arbitrario, se tendría lo mismo para el resto de los clientes.

Efectos de las covariables

En este caso se ve que todas las covariables tienen los signos esperados, tanto en presencia como en ausencia del score. Además, en los modelos (1) y (2) se observa que el efecto del score se va incrementando en valor en absoluto a medida que se avanza en los segmentos de riesgo (recordar que un segmento más alto significa menos riesgo).

Sin embargo, la antigüedad dependiente del tiempo pierde significancia en presencia del score, ya sea como regresor como variable de estratificación. Mientras que las covariables al inicio tienen efecto significativo.

Lo primero es atribuible al score dependiente del tiempo, el cual absorbe su efecto. Mientras que lo segundo podría deberse a una menor correlación con el score y la antigüedad inicial, puesto que esta última es invariante en cada cliente.

Además, si se comparan los modelos (1) y (2) se observa que tienen verosimilitudes parciales similares, pero no iguales, lo que va en contrariedad a lo que se esperaba.

Para responder ambos resultados, se comenzará con lo segundo: si bien la antigüedad es una variable predecible, existe un total de 563 clientes (5.8 % del total) que nunca abrió una cuenta corriente, por lo que su antigüedad en el tiempo siempre se mantuvo en 0. Dicho esto, se justifican las diferencias en la verosimilitud entre ambos modelos. Por lo que estrictamente no se cumpliría el ejemplo anterior, no obstante, las diferencias son pequeñas dado que son pocos los clientes que no abrieron cuenta.

Por otro lado, se observa que los modelos estratificados (3) y (4) tienen bajos valores del pseudo R^2 (esto se verá mejor si se considera el R^2 de Nagelkerke, lo cual se hará en el próximo apartado) y en los tests de significancia global, relativo a los otros modelos. Esto es debido a que estos valores se basan en la verosimilitud parcial, ignorando los efectos de la parte no paramétrica del modelo (donde alberga la información del score). Por lo tanto, se espera que los valores sean bajos. Además, se suma a ello que el resto de las covariables pierden significancia estadística al estar correlacionadas con el score, por lo que disminuye aún más estas métricas.

En relación al modelo (5), que tiene ausencia del score, también tiene un pseudo R^2 bajo, lo cual podría ser atribuible a sesgos de variable omitida.

Curvas de PI

La Figura 6.10 muestran las probabilidades de incumplimiento de algunos clientes de la cartera, utilizando los modelos (2) y (5) de la Tabla 6.6.

Lamentablemente dichas curvas no poseen las formas de una función de PI, tales como las vistas en las Figuras 6.1 y 6.5 (es decir, curvas crecientes a tasa decreciente).

En particular, algunos clientes presentaron periodos largos sin crecimiento en su PI y otros experimentaron saltos bruscos. Lo primero se debe principalmente a que los clientes perduraron largo tiempo en un segmento de bajo riesgo (1 y 2). Mientras que las alzas bruscas se explican por la caída de los clientes desde segmentos de menor riesgo a segmentos de mayor riesgo.

Sin embargo, sí se pueden generar curvas similares a las Figuras 6.1 y 6.5, pero estratificando según el score categórico dependiente del tiempo. Ello se ilustra en la Figura 6.9, la cual fue generada a partir del siguiente código:

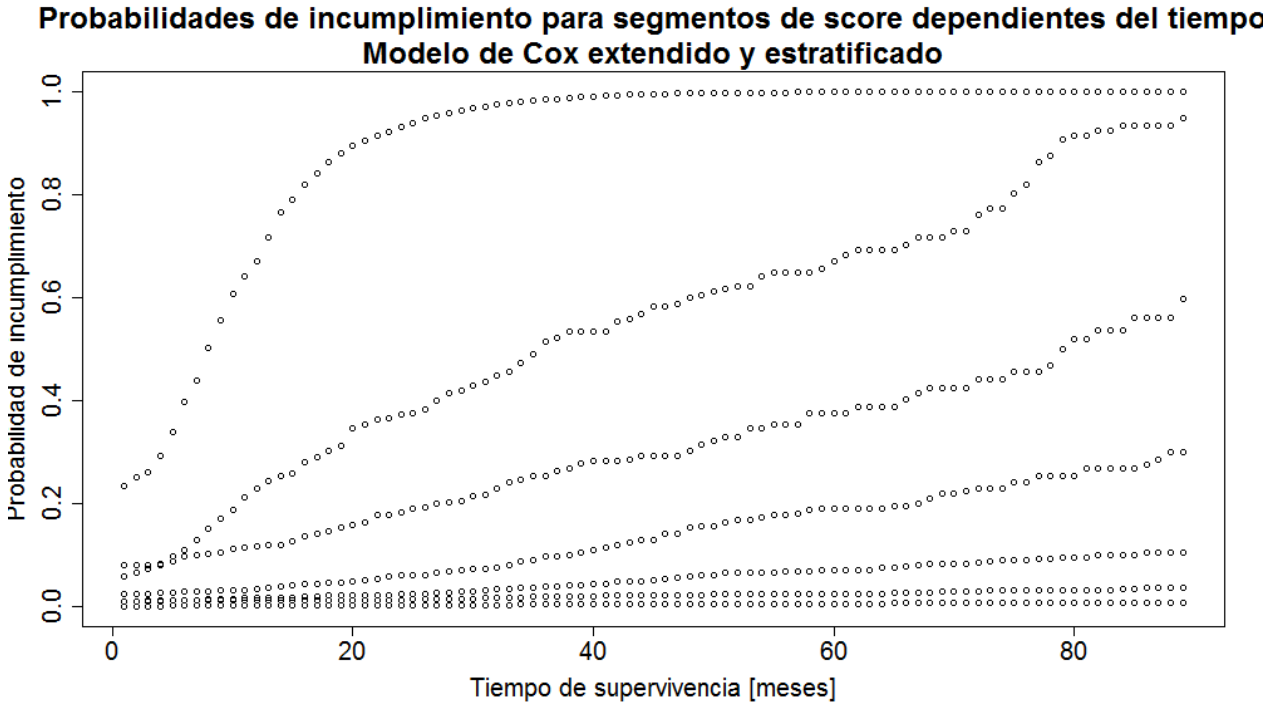
```
# ESTIMAR EL MODELO DE COX
# df_tdc_fscore = base de datos ad-hoc para modelos de Cox extendidos.

regCoxTdc <- coxph( Surv( tstart, tstop, inc ) ~ mora + antIni + reneg +
  strata( scoreCat_k1_7 ), data = df_tdc_fscore,
  model = T, x = T, y = T, ties = ‘‘Breslow" )
```

```
# ESTIMAR LA FUNCIÓN DE SUPERVIVENCIA
fit <- survfit( regCoxTdc )

# GRAFICAR LA PI ESTRATIFICADA
plot( fit$time, 1 - fit$surv )  # surv = Función de supervivencia.
```

Figura 6.9: **Probabilidades de incumplimiento crediticio para un modelo de Cox extendido y estratificado.** Se describen las curvas de PI para cada segmento o cohorte de clientes, los cuales fueron definidos de una forma dinámica, es decir, los clientes pueden variar de un segmento a otro en el tiempo. Se utilizó como referencia el modelo (5) de la Tabla 6.6



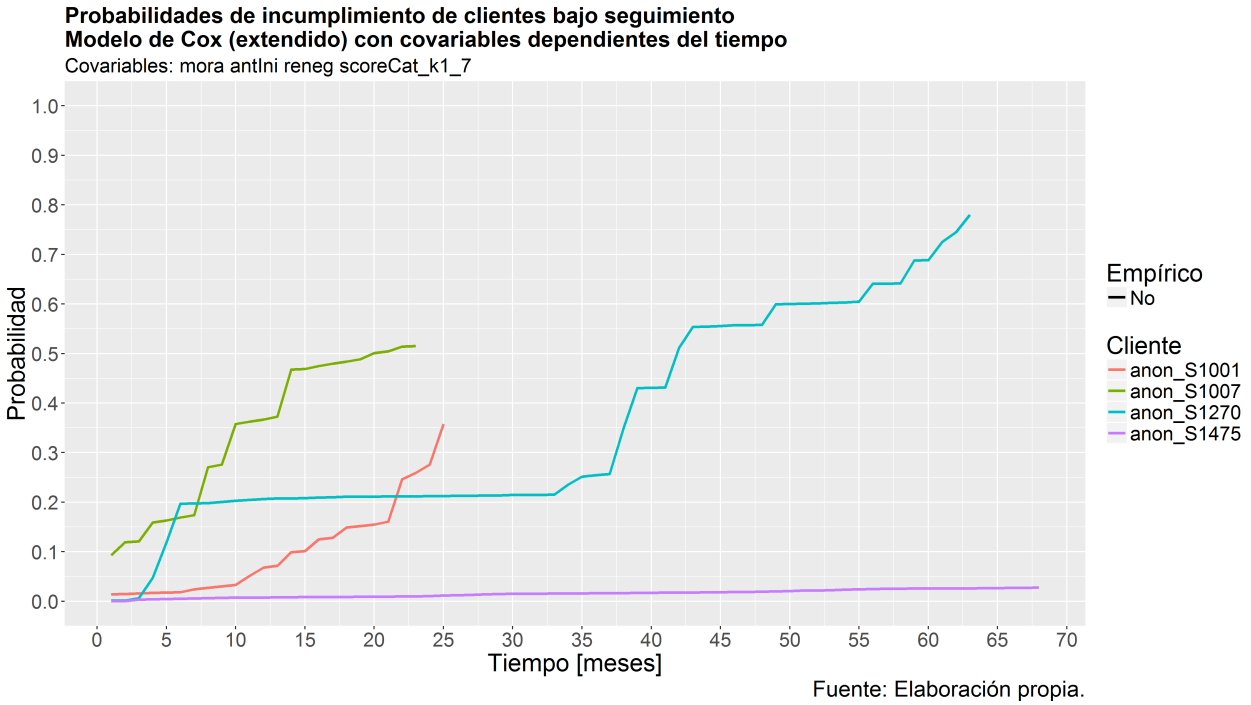
No obstante, es preciso señalar lo siguiente:

- Al aplicar el comando `survfit` sobre el objeto `coxph` (sin ninguna otra especificación), el programa detectará automáticamente el atributo `strata( )` y generará tantas curvas como estratos defina la variable dentro del atributo.
- También, en el caso de modelos con covariables (además de la variable de estratificación) el programa, por defecto, utilizará los valores promedio para estimar la supervivencia.
- Sin embargo, la documentación[2] no deja claro si para el caso especial de modelos de Cox estratos dependientes del tiempo (*time-dependent strata*), se utiliza el promedio de toda la muestra, o bien, el promedio de las covariables en el estrato. Si se lleva a cabo lo segundo, sería lo que se estaba buscando. La resolución de esta pregunta quedará como propuesta.

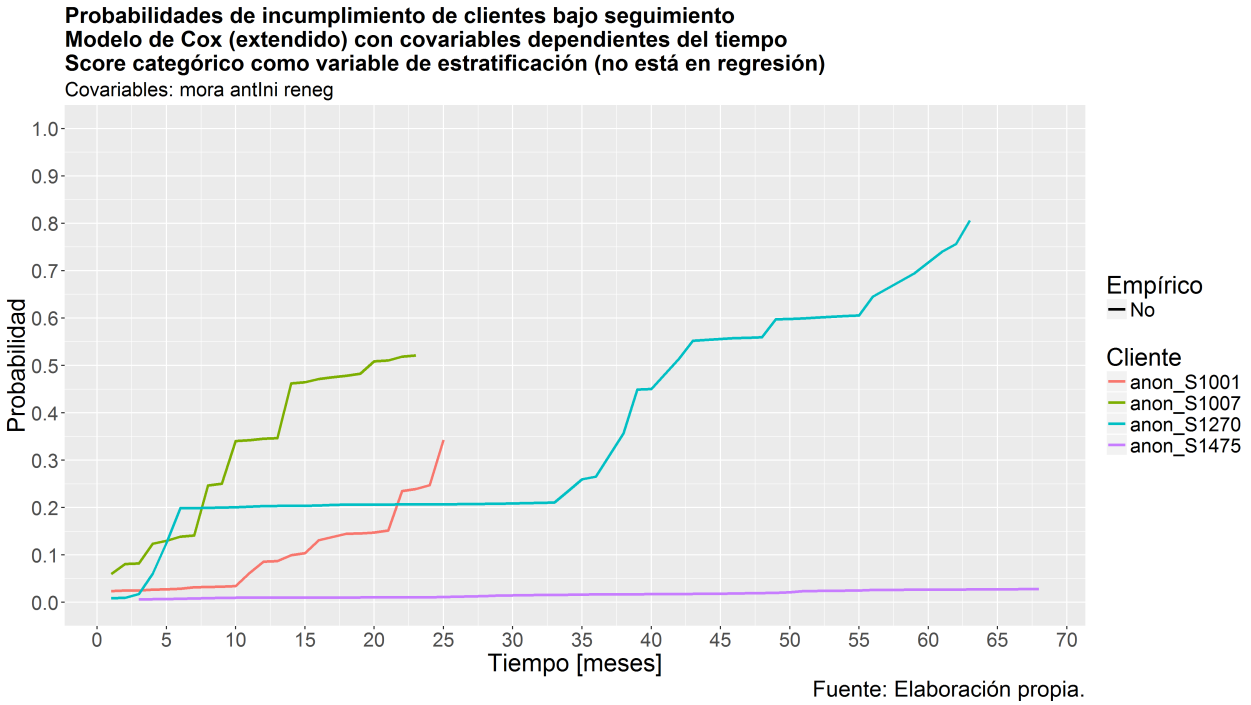
[2]Buscar la rutina *survfit.coxph* en la documentación de la librería *survival* (T. M. Therneau, 2015).

Figura 6.10: **Probabilidades de incumplimiento crediticio para algunos clientes en base modelos de Cox extendidos.** Se describen las curvas de PI para 4 clientes en base a los modelos (2) [arriba] y (5) [abajo] de la Tabla 6.6.

Glosario: **antIni** = Antigüedad inicial, **reneg** = Renegociación dependiente del tiempo, **scoreCat_k1_7** = Score categórico dependiente del tiempo, **mora** = Mora dependiente del tiempo.



Fuente: Elaboración propia.



Fuente: Elaboración propia.

6.2.3. Evaluación de modelos

En la Tabla 6.7 se reportan las métricas estadísticas para la evaluación de modelos de Cox estándar y extendidos.

En ella se observa que, según el AIC, los modelos de mejor a peor desempeño se ordenan del siguiente modo:

1. Modelos extendidos y estratificados según score categórico dependiente del tiempo.
2. Modelos extendidos con score categórico dependiente del tiempo como factor.
3. Modelos estándar estratificados según score inicial.
4. Modelo extendido en ausencia de score.
5. Modelos estándar tanto con ausencia como con presencia de score inicial categórico.

Sin embargo, en términos del pseudo R^2 los modelos estratificados, ya sean estándar o extendidos, pareciera que tuvieran una mala bondad de ajuste.

Además, pareciera que los modelos estratificados discriminan mal, puesto que sus valores del c-index rondan cerca del 50 % (es decir, la probabilidad de discriminar bien entre clientes riesgosos y no riesgosos es igual a la de obtener cara en una moneda). Lo mismo pareciera indicar la métrica iAUC para modelos de Cox estándar.

No obstante, los modelos estratificados, en términos de la verosimilitud parcial son competitivos con respecto al resto. Además, en relación a los modelos de Cox estándar, el riesgo acumulado de los residuos de Cox-Snell, para el caso estratificado, se ajustan a una distribución exponencial de tasa 1. Por lo que este tipo de modelos tienen una buena bondad de ajuste (la inclusión de la mora, genera un gráfico similar) [3].

Estas evidencias dispares tienen diferentes causas. En relación al c-index, dado se basa sólo en el riesgo relativo, es decir, en la parte que depende de las covariables; no considera los efectos del score categórico cuando es utilizado como variable de estratificación. Esto es debido a que la información del score estará contenida en la baseline hazard, la cual no es considerada en el cálculo. Esta misma explicación también se aplica para el iAUC.

Para responder la baja calidad de los modelos estratificados según los pseudo R2, se deben contemplar los resultados de la Tabla 6.8.

En ella se puede observar que los modelos estratificados poseen las últimas posiciones, si es que se ordenan según el pseudo R^2 de Nagelkerke. Es más, sus valores no superan el 0.9 %.

Sin embargo, en términos del logaritmo de la verosimilitud parcial (LL) estos modelos son competitivos respecto del resto. Incluso los modelos extendidos y estratificados son los que tienen la mayor verosimilitud.

No obstante, las versiones sin covariables de los modelos estratificados (la versiones nulas)

[3]Lamentablemente no se encontraron paquetes estadísticos que permitieran calcular los residuos de Cox-Snell para modelos con covariables dependientes del tiempo.

son las que tienen mayor verosimilitud dentro de todos los modelos. Lo que implica que la ganancia de incorporar nuevas covariables al modelo, no es significativa.

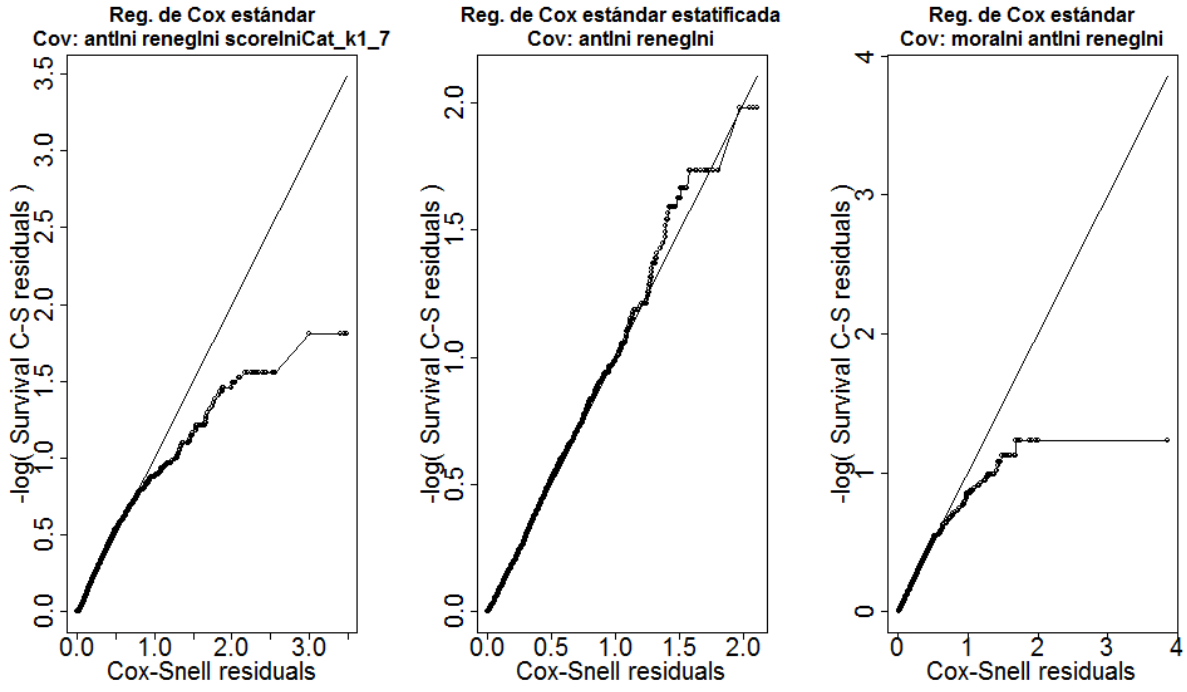
Ello se debe a que los modelos estratificados con covariables compiten contra métodos no paramétricos estratificados, mientras que los modelos sin estratificar compiten contra un método paramétrico que describe una curva única.

Si bien estos resultados van en contra de la inclusión de covariables en modelos estratificados, no son determinantes para indicar una peor bondad de ajuste respecto de los demás modelos, ya que no se está utilizando la misma vara para medir la calidad de los modelos.

Por lo tanto, no se recomienda comparar modelos estratificados y no estratificados a partir del pseudo R2 y el c-index. En cambio, se propone utilizar el AIC o el método gráfico basado en los residuos de Cox-Snell[4].

En virtud de lo anterior, las mejores alternativas corresponden a los modelos (5) y (4) de la Tabla 6.4 para los modelos de Cox estándar, y los modelos (5) y (4) de la Tabla 6.6 para los modelos de Cox extendidos.

Figura 6.11: **Residuos de Cox-Snell para modelos de Cox estándar.** La imagen muestra la función de riesgo acumulada de los residuos de Cox-Snell para los modelos (2), (4), y (5) de la Tabla 6.4. El primero corresponde a un modelo donde el score inicial categórico es regresor, el segundo lo utiliza como variable de estratificación, mientras que en el tercero el score no influye de ninguna forma. Si los puntos se acercan a la identidad, significa que el ajuste es bueno.



[4]Se espera con mucha expectativa el desarrollo de paquetes estadísticos que permitan calcular los residuos de Cox-Snell para modelos de Cox extendidos.

Figura 6.6: **Residuos de Schoenfeld para modelo de Cox con score como covariable (1).**

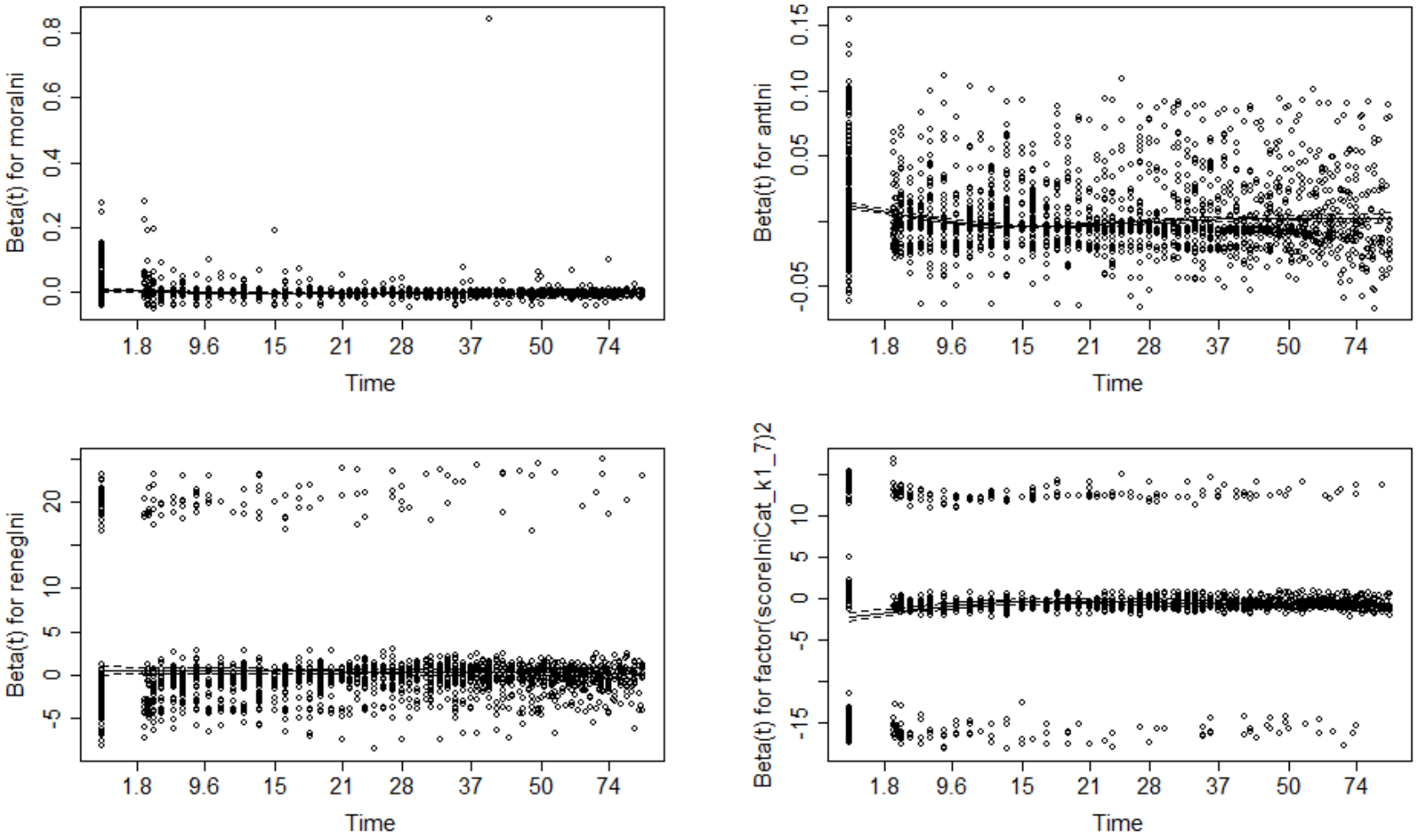


Figura 6.7: **Residuos de Schoenfeld para modelo de Cox con score como covariable (2).**

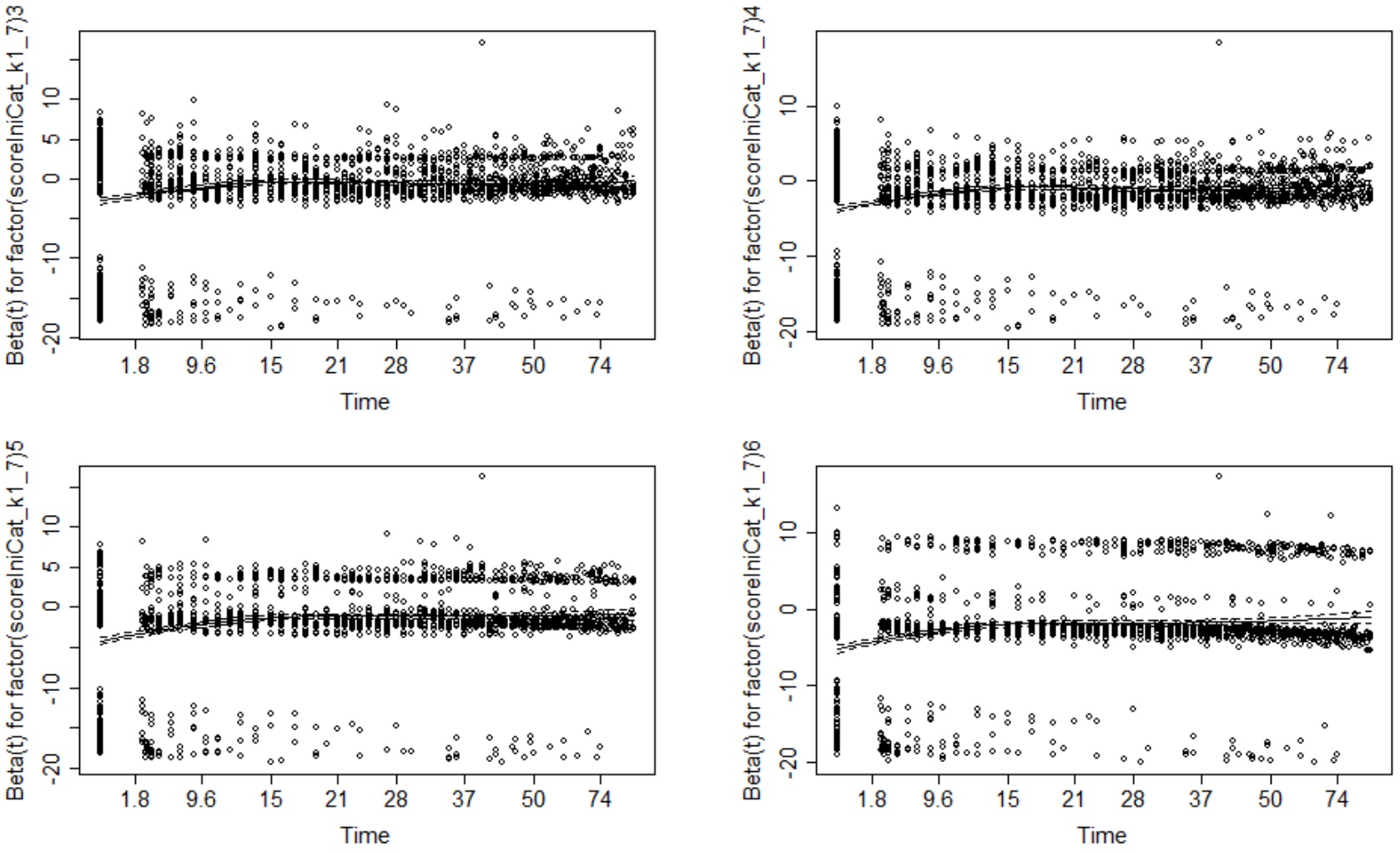


Figura 6.8: **Residuos de Schoenfeld para modelo de Cox con score como covariable (3).**

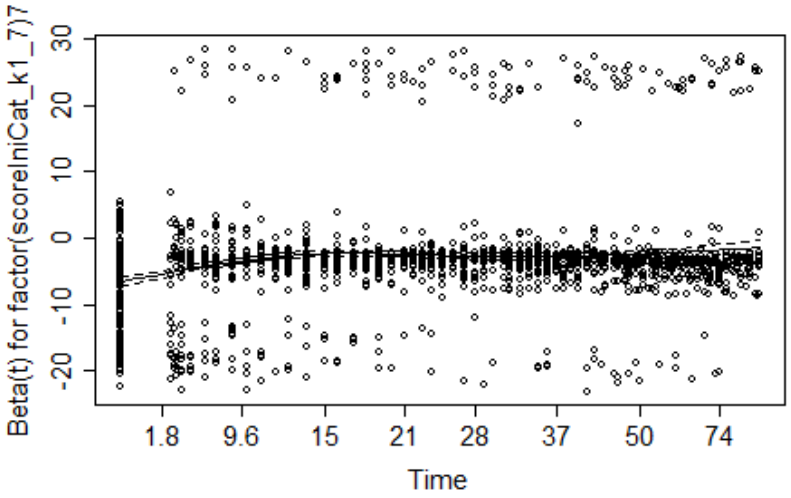


Tabla 6.7: **Métricas de evaluación estadística para modelos de Cox.** Se muestran indicadores de discriminación (CIndex y iAUC) y de bondad de ajuste (pseudo R^2 de Nagelkerke y AIC) para dos tipos de modelos de Cox: estándar y extendido. Los primeros son los reportados en la Tabla 6.4 y los segundos son los que figuran en la Tabla 6.6. Notar que el tipo de modelo (primera columna) va acompañado del número que le corresponde en la respectiva tabla de regresión. Los modelos están ordenados según menor a mayor AIC.

Glosario: **antIni** = Antigüedad inicial, **moraIni** = Mora inicial, **renegIni** = Renegociación inicial, **scoreIniCat** = Score categórico inicial, **ant** = Antigüedad dependiente del tiempo, **mora** = Mora dependiente del tiempo, **reneg** = Renegociación dependiente del tiempo, **scoreCat** = Score categórico dependiente del tiempo.

	Regresión de Cox	Covariables	Nagel-kerke R2	C-Index	iAUC	AIC	LL	g.l.
1	Extendida y estratificada (5)	`mora antIni reneg`	0.0094	0.5693		32501	-16247	3
2	Extendida y estratificada (4)	`mora ant reneg`	0.0091	0.5715		32508	-16251	3
3	Extendida (2)	`mora antIni reneg scoreCat`	0.2580	0.9337		38546	-19264	9
4	Extendida (1)	`mora ant reneg scoreCat`	0.2578	0.9341		38555	-19269	9
5	Estándar y estratificada (4)	`antIni renegIni`	0.0021	0.5248	0.5105	39554	-19775	2
6	Estándar y estratificada (3)	`moraIni antIni renegIni`	0.0021	0.5249	0.5071	39556	-19775	3
7	Extendida (3)	`mora ant reneg`	0.0307	0.7579		49198	-24596	3
8	Estándar (2)	`antIni renegIni scoreIniCat`	0.1360	0.7039	0.6897	49219	-24602	8
9	Estándar (1)	`moraIni antIni renegIni scoreIniCat`	0.1361	0.7038	0.6897	49220	-24601	9
10	Estándar (5)	`moraIni antIni renegIni`	0.0344	0.5923	0.5793	50268	-25131	3

Tabla 6.8: **Pseudos R^2 para modelos de Cox.** Se reportan el R^2 de Cox-Snell, su valor máximo (cuando la verosimilitud del modelo con covariables es 1), el R^2 de Nagelkerke (correspondiente a la división entre los dos valores anteriores), y las verosimilitudes parciales los modelos con covariables (modelo completo) y sin covariables (modelo nulo). De la tabla se concluye que las métricas basadas en los R^2 son injustas para los modelos estratificados, puesto que sus versiones nulas son mejores que aquellas provenientes de modelos no estratificados. Los modelos están ordenados según mayor a menor R^2 de Nagelkerke.

Glosario: **antIni** = Antigüedad inicial, **moraIni** = Mora inicial, **renegIni** = Renegociación inicial, **scoreIniCat** = Score categórico inicial, **ant** = Antigüedad dependiente del tiempo, **mora** = Mora dependiente del tiempo, **reneg** = Renegociación dependiente del tiempo, **scoreCat** = Score categórico dependiente del tiempo.

	Regresión de Cox	Covariables	C-S R^2	Max C-S R^2	Nagel-kerke R^2	LL	LL sin cov
1	Extendida (2)	`mora antIni reneg scoreCat`	0.0501	0.1941	0.2580	-19264	-25284
2	Extendida (1)	`mora ant reneg scoreCat`	0.0500	0.1941	0.2578	-19269	-25284
3	Estándar (1)	`moraIni antIni renegIni scoreIniCat`	0.1354	0.9949	0.1361	-24601	-25297
4	Estándar (2)	`antIni renegIni scoreIniCat`	0.1353	0.9949	0.1360	-24602	-25297
5	Estándar (5)	`moraIni antIni renegIni`	0.0342	0.9949	0.0344	-25131	-25297
6	Extendida (3)	`mora ant reneg`	0.0060	0.1942	0.0307	-24596	-25297
7	Extendida y estratificada (5)	`mora antIni reneg`	0.0012	0.1305	0.0094	-16247	-16391
8	Extendida y estratificada (4)	`mora ant reneg`	0.0012	0.1305	0.0091	-16251	-16391
9	Estándar y estratificada (3)	`moraIni antIni renegIni`	0.0021	0.9840	0.0021	-19775	-19785
10	Estándar y estratificada (4)	`antIni renegIni`	0.0021	0.9840	0.0021	-19775	-19785

Capítulo 7

Selección de modelos

En el Capítulo 6 se determinaron los mejores modelos paramétricos y semi-paramétricos, desde un punto de vista estadístico.

En este apartado se definirán los mejores modelos de acuerdo a los posibles usos en la empresa, entre ellos destaca su uso como modelo de originación o como modelo de seguimiento.

Cabe recordar que los modelos de orginación son aquellos que tienen por objeto ver la probabilidad de incumplimiento de clientes que recién están entrando en una relación con la entidad bancaria. Mientras que los modelos de seguimiento buscan evaluar la PI de clientes antiguos conforme pasa el tiempo.

Modelos de originación

Se postula que los modelos con covariables al inicio son adecuados para modelos de originación, ya que la evaluación de un nuevo cliente se hará por medio de un modelo calibrado con individuos que compartieron su condición. Por lo tanto, los modelos más adecuados para su uso son la regresión AFT lognormal con score categórico inicial como regresor (modelo (4) de la Tabla 6.1) y el modelo de Cox estándar estratificado según score categórico (modelo (5) de la Tabla 6.4).

Sin embargo, entre ellos existen ventajas y desventajas. En efecto, el modelo semi-paramétrico mostró tener una muy buena bondad de ajuste, debido a que los residuos de Cox-Snell distribuyen aproximadamente como una exponencial de tasa 1 (ver Figura 6.11).

En cambio, la regresión AFT lognormal, si bien fue la que mejor ajustó en términos de los residuos de Cox-Snell, presentó observaciones con desviaciones importantes respecto de la identidad (ver Figura 6.4).

Por otro lado, el modelo paramétrico permite realizar pronósticos para tiempos arbitrarios, puesto que la fórmula de la PI depende explícitamente del tiempo. Por lo tanto, al momento de analizar un nuevo cliente, solamente será necesario recopilar su información inicial, para

luego realizar proyecciones en los tiempos que el analista estime conveniente.

Mientras que la estimación semi-paramétrica, además de necesitar la información inicial del cliente, requiere de una estimación no paramétrica de PI basal (i.e. $1 - \hat{S}_0(t)$), la cual, dependiendo del tipo de modelo, requerirá de todos los clientes de la muestra de entrenamiento, o bien, de aquellos pertenecientes al estrato que le corresponde al individuo evaluado. Por lo tanto, las predicciones de la PI estarán condicionadas al máximo tiempo de falla observado en en la muestra o en el estrato, respectivamente (verificar en la Figura 6.5).

Por lo tanto, a pesar de la menor bondad ajuste, un modelo paramétrico es mejor para hacer pronósticos. Por lo que la alternativa elegida sería el modelo (5) de la Tabla 6.1.

Modelos de seguimiento

Para los modelos de seguimiento se recomienda utilizar modelos de Cox extendidos y estratificados, es decir, los modelos (5) y (4) de la Tabla 6.6. Los que también resultaron ser los de mejor calidad, en términos del AIC (ver 6.7).

Sin embargo, esta proposición debe tomarse con cautela, puesto que no se recomienda describir curvas para clientes únicos bajo este tipo de modelos. En cambio, sí se recomienda que se utilicen curvas para diferentes cohortes o segmentos de clientes, pues recrean trayectorias que hacen más sentido. No obstante, es preciso tener claridad en torno al cálculo empleado por la rutina que define las PI estimadas.

Por otro lado, se tiene que tener presente que este tipo de modelos, al igual que todos los semi-paramétricos, no pueden realizar pronósticos para tiempos arbitrarios. Por lo que se propone investigar un enfoque paramétrico para el modelamiento de la PI con covariables dependientes del tiempo.

Finalmente se recomienda la búsqueda de paquetes estadísticos que permitan calcular los residuos de Cox-Snell para modelos con covariables dependientes del tiempo, como también, la búsqueda de nuevos métodos para evaluar calidad de los modelos.

Comentarios adicionales

Para finalizar el capítulo se harán recomendaciones adicionales, en torno a las variables a utilizar en los modelos. La condición de resumir el riesgo del cliente en una sola variable, hace que el score sea muy versátil para definir distintos cohortes o segmentos de clientes.

Sin embargo, su incorporación como insumo para el pronóstico de las PI en el tiempo, es problemática, puesto que presenta alta correlación con otras covariables y porque nace de un modelo que ya busca predecir una probabilidad de incumplimiento (aunque puntual y de 1 a 12 meses en el futuro).

Si no se desea desechar el score para el pronóstico, la única alternativa aparente sería

ajustar modelos diferentes en cada submuestra inducida por los tramos de score. No obstante, ello introduce la complejidad de ajustar tantos modelos como segmentos de clientes haya.

De lo contrario, si se desea ajustar estas PI a través de un solo modelo, y si además se busca copiar la versatilidad del score como variable de segmentación; se recomienda crear un indicador distinto que permita resumir el riesgo del cliente el cual pueda usarse tanto como regresor o como variable de estratificación en los modelos.

Capítulo 8

Conclusiones

Conclusiones generales y selección de modelos

Se concluye que el incumplimiento crediticio se puede ver como un fenómeno de supervivencia, dada la estructura de datos que presenta.

Además, este enfoque también permite estimar una PI temporal según segmentos o cohortes de clientes, lo cual se puede hacer por medio de una variable de clasificación incorporada como regresor o como variable de estratificación (lo último en modelos de Cox).

En términos generales, para el pronóstico de las PI, se recomiendan los modelos paramétricos, ya que definen una función que depende exclusivamente del tiempo. Ello permite predecir probabilidades para horizontes arbitrarios.

Sin embargo, debido a limitaciones del paquete estadístico, no se estimaron PIs paramétricas con covariables dependientes del tiempo. Por lo que queda propuesta la búsqueda de librerías que permitan realizar ello.

Es por lo anterior que para modelos de originación se recomienda el uso de una regresión AFT lognormal evaluada con covariables al inicio. Esto es debido a que nuevos clientes pueden ser evaluados a partir de un modelo entrenado con individuos que compartieron su condición. En términos estadísticos, los residuos de Cox-Snell de este modelo son los que mejor se ajustaron a una distribución exponencial de tasa 1, dentro de todos los modelos paramétricos, lo que justifica su elección.

Mientras que para modelos de seguimiento se propone utilizar la regresión de Cox con covariables dependientes del tiempo, donde el score se utilice como variable de estratificación. Se recomienda su uso, puesto que incorpora la historia del cliente, tanto en el entrenamiento como en el pronóstico de las curvas. En términos estadísticos, fue el modelo con menor AIC dentro de todos los modelos semi-paramétricos.

No obstante, para este último no se recomienda hace pronósticos para clientes individuales, pues las curvas respectivas no describen trayectorias bien comportadas (crecientes a tasa

decreciente). En cambio, se propone describir las curvas según segmento.

Un fenómeno importante en los estudios de supervivencia es la censura (derecha), la cual influye en la expresión general de la función de verorsimilitud (ver Kalbfleisch y Prentice (2002), Klein y Moeschberger (2005) y Collett (2015)). Además, provoca que la función de distribución empírica no refleje la realidad de los datos. Por lo tanto, medidas de calidad de pronóstico basadas en la distancia de la PI predicha contra la función empírica, no son adecuadas.

Es por ello que se recomiendan criterios de calidad que se sustenten en resultados teóricos, entre los que destacan los residuos de Cox-Snell y métricas tales como el AIC, los pseudo R2 y el C-index.

Efectos de las covariables

Se demostró que la mora y la renegociación tienen un efecto que favorece el incumplimiento, mientras que la antigüedad y el score tienen un efecto contrario. En particular, el score y la renegociación son las covariables con mayores efectos.

No obstante, la inclusión del score ya sea como regresor o como variable de estratificación, absorbe el efecto de las otras covariables.

Además, como su construcción se fundamenta en las PI predichas por un modelo de credit scoring subyacente, se conjetura un problema de causalidad reversa.

Para evitar esos problemas existen dos opciones: ajustar modelos independientes sobre cada submuestra inducida por el score (siendo costoso en tiempo), o crear un nuevo indicador que resuma la información del cliente y que no derive de un modelo que prediga una probabilidad, de modo que pueda utilizarse como regresor o como variable de estratificación.

Capítulo 9

Proyecciones y trabajos futuros

Múltiples tipos de eventos

Dado que el panel inicial sólo distiguía al incumplimiento como único tipo de evento, se tuvo que suponer censura derecha para el resto de los clientes que salieron de la base.

Este puede ser un supuesto no realista por dos razones: (1) las personas que salieron puede que hayan pagado todas sus obligaciones, y (2) es poco probable que haya clientes a los que se les pierda el rastro. De cumplirse (1) se tendría la existencia de otro evento: el pago. SI se cumple (2) en su totalidad, no habría censura derecha.

Este último aspecto sería una buena noticia, por cuanto sería factible implementar criterios de calidad de pronóstico en base a la función de distribución empírica.

Sin embargo, la existencia de dos tipos de eventos, requerirá extender los modelos de supervivencia vistos. Una posibilidad es utilizar el enfoque de *competing risks*, el cual consiste en hacer competir los tiempos teóricos de falla y de pago en un mismo modelo, para luego ver cuál se materializará primero.

En términos matemáticos, el tiempo observable de supervivencia debería estar dado por:

$$T_{\mathrm{i}} = \mathrm{mín}\{T_{\mathrm{i}}^{\mathrm{incump}}, T_{\mathrm{i}}^{\mathrm{pago}}, C_{\mathrm{i}}\} \tag{9.1}$$

Donde $T_{\mathrm{i}}^{\mathrm{incump}}$ es el tiempo teórico de incumplimiento, $T_{\mathrm{i}}^{\mathrm{pago}}$ es el tiempo teórico de pago y C_{i} es el tiempo teórico de censura del cliente i.

Para su estimación se recomienda ver los trabajos de Bravo, Thomas, y Weber (2012), Banasik y cols. (1999) y Stepanova y Thomas (2001). Cabe indicar que el paquete `survival` también tiene la opción de estimar modelos bajo el enfoque de *competing risks*.

Incorporar información macroeconómica

La incorporación de información macroeconómica en los modelos, además de ser una forma de buscar mejor calibración, permitiría describir los efectos macroeconómicos sobre distintos tipos de clientes.

Estos efectos se pueden incluir por medio de interacciones con covariables asociadas a los clientes, tal como se hace en el trabajo de Bellotti y Crook (2009). Ello posibilitaría también la realización de pruebas de tensión sobre clientes, al alterar las condiciones macroeconómicas.

Estabilidad a largo plazo

Bogren (2015) y Tong, Mues, y Thomas (2012) dieron cuenta de la existencia de un segmento de usuarios de crédito que nunca incumplen. Es por ello que para recrear dicho fenómeno aplicaron el enfoque de *mixture cure models*, el cual plantea la existencia de dos clases latentes: los clientes que incumplen y aquellos que no.

La idea de estos modelos reside en verificar a qué segmento pertenece el individuo y, condicional a ello, observar cuánto tiempo demora en incumplir. La probabilidad de pertenecer a un segmento se denomina como *incidencia*, mientras que la función de supervivencia condicional a estar en dicho segmento, se define como *latencia*. Tanto la latencia como la incidencia deben considerar variables que describan atributos del cliente, aunque no necesariamente tienen que ser las mismas (Bravo y cols., 2012).

Se conjetura que al incorporar dos clases latentes se tendría como consecuencia que las curvas de supervivencia converjan en el largo plazo.

Otra forma es utilizar una analogía a la idea planteada por Rodríguez (2016), la cual consiste en calcular tasa de supervivencia de los clientes que nunca van a incumplir $S(\infty)$ y luego aplicar los modelos de supervivencia antes vistos para quienes eventualmente van a incumplir, lo cual se resumirá en la función $S(t)$.

Para este último tipo de clientes, la variable aleatoria del tiempo de falla T estará bien definida, puesto que en el infinito todos terminarán incumpliendo, por lo tanto, la integral de la densidad a lo largo de todos los tiempos de falla, daría 1.

Luego de ello, se definen nuevas densidades, funciones de supervivencia y funciones de riesgo corregidas (condicionadas) por los clientes que nunca incumplirán:

$$f^*(t) = \frac{f(t)}{1 - S(\infty)} \qquad S^*(t) = \frac{S(t) - S(\infty)}{1 - S(\infty)} \qquad h^*(t) = \frac{f^*(t)}{S^*(t)} = \frac{f(t)}{S(t) - S(\infty)}$$

Covariables internas y externas

En el contexto de los modelos de Cox extendidos, Kalbfleisch y Prentice (2002) propusieron distinguir dos tipos de covariables: las internas y las externas. Las internas (o endógenas) son aquellas que se originan en el mismo individuo y que sus futuras realizaciones (*future path*) dependen de la vida del cliente. Es decir, si el cliente incumple y, en consecuencia, es eliminado de la base, entonces no existirán futuras realizaciones de la variable.

En cambio, las covariables externas (o exógenas) son aquellas que emergen de un proceso independiente al cliente. Por ejemplo, se puede pensar en los efectos del smog sobre los pacientes de asma, o bien, en los efectos de shocks macroeconómicos sobre el comportamiento de pago de los clientes.

Esta acotación es relevante, puesto que si la covariable es interna, ésta no será predecible en el sentido de Kalbfleisch y Prentice (sus futuros valores dependenderán del tiempo de falla) y, por lo tanto, la función $S(t, X(t))$ ya no tendría una interpretación como probabilidad de supervivencia. No obstante, este aspecto aún no ha sido solucionado en la literatura.

Sin embargo, una alternativa busca abordar este problema: los *joint models* (Rizopoulos, 2012). Los cuales, además de posibilitar la inclusión de información longitudinal (covariables/marcadores dependientes del tiempo), fueron desarrollados bajo la promesa de lidiar con el problema de las covariables internas.

Lidiando con la causalidad revesa

Según T. Therneau y cols. (2016), existe la posibilidad de que ciertas covariables pueden tener causalidad reversa con respecto a los tiempos de falla. Por ejemplo, si se define una covariable que indique la última visita de los familiares de un paciente terminal.

Este caso cobra especial relevancia en el problema, puesto que de la exploración de datos se determinó que valores de las variables cercanos al tiempo de falla, podrían ser indicativos del incumplimiento. Entonces cabría preguntarse si el valor de las covariables implican el incumplimiento, o bien, si cumplimiento inminente determina el valor de éstas.

Para ello, Therneau implementó la opción delay sobre la función coxph (paquete *survival* de R), el cual permite estimar el modelo de Cox con covariables rezagadas. Por ejemplo, se podría decir que valores del score 3 meses antes podría ser indicativo del default en el tiempo actual.

Capítulo 10

Anexo 1: Detalles sobre IFRS9

El IFRS9, publicado en Julio de 2014 por el Buró Internacional de Estándares Contables (IASB), es un documento que actualiza las prácticas en el reporte de los estados financieros, el cual será aplicable desde Enero de 2018 en todas las entidades bancarias en Europa.

Este estándar se sustenta en tres pilares fundamentales: clasificación y medición, deterioro crediticio y contabilidad de la cobertura. Cada uno de ellos busca respectivamente:

- Simplificar los requerimientos para la clasificación y medición de los activos y pasivos en los estados financieros.
- Tener un reconocimiento temprano de las pérdidas de crédito (PE), relacionando de mejor forma las pérdidas con el tiempo, a través de un esquema único de deterioro crediticio.
- Vincular de mejor forma la gestión de riesgos y el manejo contable.

En relación al segundo ámbito, el estándar incorpora un nuevo esquema que considera tres etapas según el nivel de riesgo del préstamo. En cada una de estas etapas se proponen diferentes formas de calcular la PE, influyendo también en el cálculo de la PI (que es el aspecto de interés para esta memoria).

Las etapas son las siguientes:

Etapa I: Consiste en el reconocimiento inicial del riesgo del instrumento, lo cual, según el estándar, debe hacerse por medio de métodos que sean simples y no costosos en esfuerzo.

Para todos estos nuevos préstamos, la PE se debe calcular a través a través del producto entre el efecto total de la pérdida del crédito y la probabilidad que ésta se materialice dentro de los próximos 12 meses.

Es decir, se debe emplear la fórmula convencionalmente extendida en el mundo bancario para el cálculo de la pérdida esperada: $PE = EI \times PI \times PDI$ (Basel Committee On Banking Supervision, 2005). Donde EI corresponde a la exposición total al incumplimiento; PI, a la probabilidad de incumplimiento entre 1 a 12 meses; y PDI, a la pérdida dado el incumplimiento.

Notar que el efecto total de la pérdida del crédito equivale a $EI \times PDI$.

Etapa II: Cuando existe un aumento significativo en el riesgo (se recomienda un método simple para verificarlo), el préstamo pasa a una nueva etapa, en la cual se reconoce el valor de vida de las pérdidas esperadas (LECL) como indicador de la PE.

Según el estándar, el LECL corresponde a la esperanza del valor presente de las pérdidas de crédito[1] que surgirían si un cliente incumple a lo largo de la vida del instrumento. Dicho de otro modo, es un promedio ponderado de las pérdidas donde la PI es el peso.

Etapa III: Si el riesgo llega a un nivel en el cual el préstamo se declara como deteriorado, éste pasa a la última etapa, en la que también se reconoce al LECL como indicador de la pérdida esperada.

En relación a la PI, el estándar da a entender que es un factor único que representa la probabilidad de incumplir a lo largo de la vida del crédito, es decir, una PI de largo plazo o *lifetime*. La cual va más allá del límite de los 12 meses que establece la PI convencional.

Dado lo anterior, resulta necesario un método o modelo que permita describir estas probabilidades para duraciones superiores a los 12 meses.

10.1. Discusión entre el sector contable y regulatorio

No obstante, el nuevo estándar ha traído consigo reacciones de parte del Comité de Basilea, el cual emitió una Guía para la Contabilización de las Pérdidas Esperadas de Crédito, la cual dedicó un anexo completo a IFRS9. En él, si bien existen concordancias, hay una diferencia sustantiva en lo que respecta al cálculo de las pérdidas esperadas (Basel Committee On Banking Supervision, 2015):

> *"El Comité enfatiza que un monto igual a la pérdida esperada a 12 meses no es sólo la pérdida esperada en los próximos 12 meses, sino que son los descalces a lo largo de la vida del instrumento, debido a eventos de pérdida a materializarse en los próximos 12 meses."*

Es más, el Comité de Basilea también propone una metodología que, a través de un factor de corrección sobre la fórmula de Basilea II, realiza ajustes por la madurez de los instrumentos (Basel Committee On Banking Supervision, 2005).

A pesar de esta diferencia entre las entidades regulatorias y contables, de la cual se espera una pronta convergencia en el futuro, es preciso que la entidad bancaria recopile mayores antecedentes para informarse, o bien, tomar parte en esta discusión, de modo de adaptarse

[1]Una denominación más precisa que pérdida de crédito sería la de *descalce de caja* o *cash shortfall*, que corresponde a la diferencia entre el flujo que se espera recibir y lo que efectivamente se recibe en un tiempo determinado. Este descalce puede surgir tanto de un atraso en los pagos (solucionable en el corto plazo) como de una pérdida efectiva (cuando ya no se espera el pago). En este apartado, este término y el de pérdida se tratarán de forma indistinta.

al contexto de los próximos años, independiente de qué posición domine.

De este modo, el desarrollo de un nuevo método para calcular las PI en tiempo, permitiría que la entidad pueda anticiparse al futuro escenario.

Capítulo 11

Anexo 2: Modelos de scoring basados en regresión logística

Este tipo de modelos buscan predecir el comportamiento de la variable dependiente binaria que toma el valor de 1 cuando se declara un incumplimiento entre 1 a 12 meses más y 0 si no. Para lo anterior se corre un modelo de regresión logística, que entrega como output una probabilidad teórica que cumple con la siguiente ley:

$$\hat{p_{\mathrm{i}}} = \frac{\mathrm{e}^{X_{\mathrm{i}}'\hat{\beta}}}{1 + \mathrm{e}^{X_{\mathrm{i}}'\hat{\beta}}} \tag{11.1}$$

Es decir, la probabilidad tiene una relación no lineal con las covariables.

Una práctica usual en el sector bancario (de la cual no se exime la entidad) es aplicar una transformación sobre la probabilidad teórica para:

- Obtener una relación lineal con las covariables.
- Que sea de fácil interpretación.
- Que permita ordenar a los clientes según su capacidad de pago y, de esta forma, definir criterios para diferenciar entre buenos y malos pagadores.

Así se llega al concepto de *score*, el cual se obtiene mediante la siguiente transformación sobre la probabilidad teórica (Siddiqi, 2006):

$$\begin{aligned} \text{score} &= \ln(\text{odds}) \times \text{factor} + \text{offset} \\ &= \sum_{j,\mathrm{i}=1}^{k,n} \left(- \left(\text{WOE}_j \beta_{\mathrm{i}} + \frac{a}{n} \right) \times \text{factor} + \frac{\text{offset}}{n} \right) \end{aligned}$$

Donde

$$\begin{aligned} \text{WOE}_j &= \text{weight of evidence de cada atributo} \\ \beta_{\text{i}} &= \text{coeficiente de regresión para cada característica} \\ a &= \text{intercepto de la regresión logística} \\ n &= \text{número de características} \\ k &= \text{número de atributos en cada característica} \\ \ln(\text{odds}) &= \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = X'\beta \end{aligned}$$

No obstante, para otros aspectos del quehacer bancario se requiere volver al concepto de probabilidad. En particular, para calcular las pérdidas esperadas de crédito, las cuales tienen impacto sobre el capital regulatorio y las políticas de *pricing*.

Capítulo 12

Anexo 3: Apéndice teórico en análisis de supervivencia

12.1. Profundización en modelos AFT paramétricos

A continuación se deducirá la función de supervivencia de un individuo de covariables Z, según el modelo de tiempo de falla acelerado (AFT).

Supóngase que la relación del tiempo de falla y la información del cliente está dada por:

$$\ln T \equiv Y = \alpha + Z'\beta + \sigma W \tag{12.1}$$

Donde α es el intercepto, β son los coeficientes asociados a las columnas de la matriz de covariables Z (fijas), W es el término de error de la regresión y σ es un parámetro que amplifica el error.

Notar que W tiene distribución F_W que tiene media 0.

Si se aplica la función exponencial en ambos lados de la ecuación, se tiene:

$$T = \mathrm{e}^{\alpha+\sigma W}\mathrm{e}^{Z'\beta} = T_0 \cdot \mathrm{e}^{Z'\beta} \tag{12.2}$$

Donde T_0 es la variable aleatoria asociada al tiempo de falla del individuo de covariables $Z = 0$.

Por otro lado, se tiene que la función de supervivencia de una variable aleatoria T condicional a la información Z, está dada por:

$$S(t, Z) = \Pr(T > t|Z) \tag{12.3}$$

Utilizando 12.2 se tendrá que:

$$S(t, Z) = \Pr(T > t|Z) = \Pr(T_0\mathrm{e}^{Z'\beta} > t) = \Pr(T_0 > t\mathrm{e}^{-Z'\beta}) = S_0(t\mathrm{e}^{-Z'\beta}) \tag{12.4}$$

Lo que cumple con la expresión dada en el Capítulo 3.

Además se pueden definir las funciones de densidad de probabilidad y de riesgo como se sigue:

$$f(t,Z) = -\frac{\mathrm{d}}{\mathrm{d}t}S(t,Z) = -\frac{\mathrm{d}}{\mathrm{d}t}S_0(t\mathrm{e}^{-Z'\beta}) = f_0(t\mathrm{e}^{-Z'\beta})t\mathrm{e}^{-Z'\beta} \tag{12.5}$$

$$h(t,Z) = \frac{f(t,Z)}{S(t,Z)} = \frac{f_0(t\mathrm{e}^{-Z'\beta})t\mathrm{e}^{-Z'\beta}}{S_0(t\mathrm{e}^{-Z'\beta})} = h_0(t\mathrm{e}^{-Z'\beta})t\mathrm{e}^{-Z'\beta} \tag{12.6}$$

Notar que tanto la distribución de T como la de T_0 se definen a partir de la distribución de W. En particular se definirá la distribución de T_0 (la de T es directa de ella):

$$\Pr(T_0 > t) = \Pr\left(\frac{\ln T_0 - \alpha}{\sigma} > \frac{\ln t - \alpha}{\sigma}\right) \tag{12.7}$$

Si se utiliza la parametrización $\lambda = \exp(-\alpha)$ y $\gamma = 1/\sigma$, se obtiene:

$$\Pr(T_0 > t) = \Pr(W > \gamma \ln(\lambda t)) \tag{12.8}$$

Con ello se define completamente la distribución del tiempo de falla.

En particular, si se supone que $\sigma = 1$ y que $F_W(w) = 1 - \exp(-\mathrm{e}^w)$ ($W \sim$ valor extremo), la función de supervivencia de T_0 estará dada por:

$$S_0(t) = \exp(-\mathrm{e}^{\ln(\lambda t)}) = \exp(\lambda t) \tag{12.9}$$

Lo que equivale a la función de supervivencia de una distribución exponencial de tasa λ.

Finalmente la función de supervivencia para T estará dada por:

$$S(t,Z) = \exp(\lambda \mathrm{e}^{-Z'\beta}t) \tag{12.10}$$

De la misma forma se pueden se construir distribuciones para el tiempo de falla, suponiendo otras distribuciones. En particular, si $\sigma > 1$ y que $W \sim EVT1$, se tendrá una distribución Weibull para T. En cambio, si $W \sim N(0,1)$, T será lognormal. Lo cual se dejará propuesto.

12.2. Test para supuesto de riesgos proporcionales en modelos de Cox

Para evaluar la proporcionalidad del riesgo sobre una variable dicotómica, es usual graficar las curvas de riesgo acumulado y las funciones de supervivencia para las distintas submuestras inducidas por dicha variable: si la distancia entre las curvas se mantiene constante, se determina que se cumple el supuesto de proporcionalidad.

No obstante, para el caso de las variables continuas (como lo es en el caso de estudio), la verificación es más compleja. Para estos casos -aunque también extensible para *dummies* y categóricas- Grambsch y Therneau (1994) propusieron un test para la proporcionalidad que busca verificar si los efectos de las covariables son constantes en el tiempo. Para ello se valieron del siguiente resultado:

Dados $0 < t_1 < \ldots < t_k < t_{\mathrm{d}}$ los diferentes tiempos de eventos, donde d es el total de eventos, y s_k^* el vector de dimensión $p \times 1$ de los residuos de Schoenfeld escalados para el evento k. Si $\hat{\beta}_{p\times 1}$ son los parámetros estimados por un modelo de Cox ordinario (i.e. con variables al inicio de la medición), se tiene que:

$$E(s_{jk}^*) + \hat{\beta}_j \approx \beta_j(t_k) \tag{12.11}$$

Es decir, que el parámetro verdadero de la covariable j como función del tiempo, depende de la evolución de los residuos a obtener por el modelo. Esto permite elaborar un gráfico de dispersión entre los residuos y el tiempo, tal como se ilustra en la figura 12.1[1].

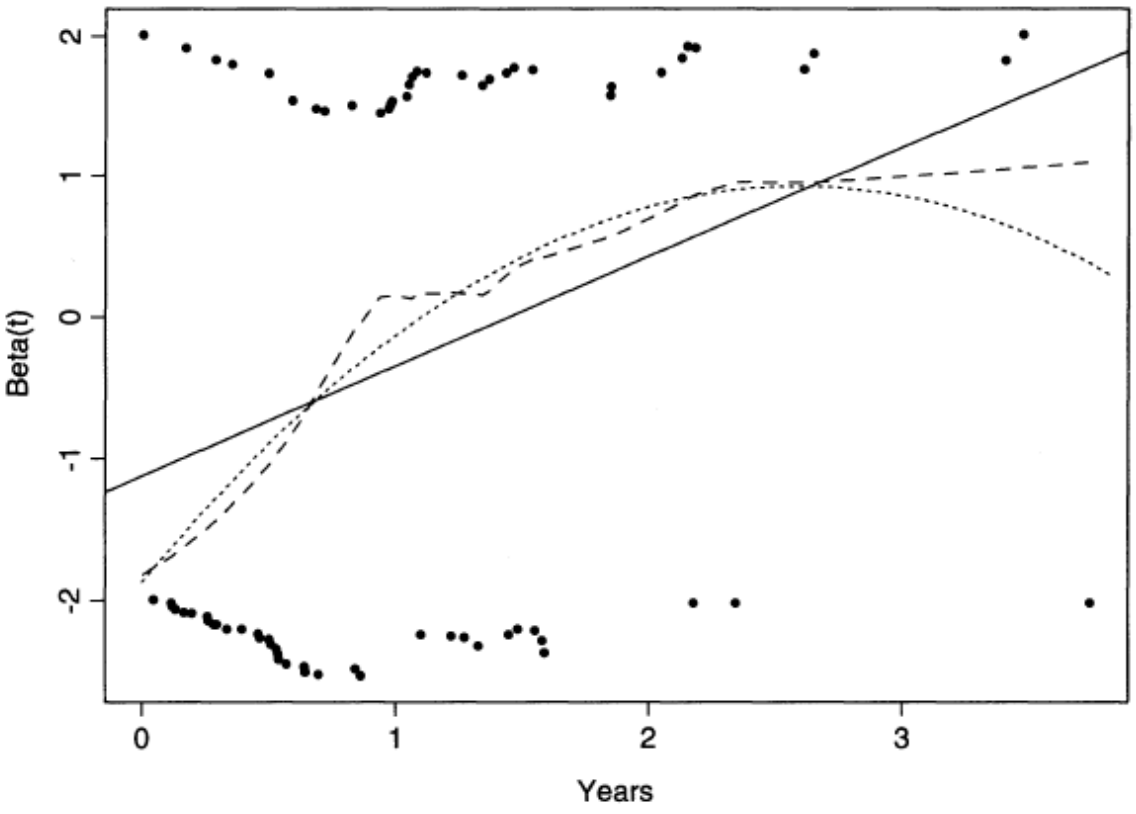


Figura 12.1: **Residuos de Schoenfeld en función al tiempo.** Fuente: T. M. Therneau y Grambsch (2000)

Así, la hipótesis nula del test de proporcionalidad corresponde a $\beta_j(t) = \beta_j \equiv cte$, mientras

[1]Obsérvese que en la Figura 12.1 se ven curvas que describen ajustes lineales, cuadráticos y un suavizamiento *lowess* implementado en diferentes paquetes estadísticos. La idea es determinar por inspección qué forma funcional es *ad-hoc* para g

que el estadístico del test nace del planteamiento de un modelo lineal generalizado entre $\beta_j(t)$ y una función del tiempo $g(t)$, la cual busca ajustar curvas sobre los datos dispersos.

En efecto, se plantea:

$$\beta_j(t) = \beta_j + \theta_j(g_j(t) - \bar{g}_j), j \in 1 \dots p \tag{12.12}$$

Donde la hipótesis nula corresponde a $H_0 : \theta = 0$, la cual se contrasta con una distribución χ^2_p. Mientras que la hipótesis para cada covariable se testea con una χ^2_1.

Así, en caso de violarse el supuesto de proporcionalidad de los riesgos, se pueden implementar modelos estratificados o modelos con variables dependientes del tiempo, los cuales se describirán en los apartados siguientes.

Tanto el test como los gráficos de dispersión están implementados en la función *cox.zph* del paquete *survival* de R. Los cálculos de los residuos de Schoenfeld están implícitos en la misma rutina, no obstante, también se pueden obtener a través de la función *predict.coxph*.

12.3. Profundización en la métrica iAUC

Una de las formas de evaluar el poder discriminativo de un modelo de supervivencia es a través del enfoque ROC/AUC dependientes del tiempo (Heagerty y Zheng, 2005; Saha y Heagerty, 2010).

Para esta formulación curva ROC está definida por las medidas de *sensitivity* y *specifity*, las cuales están dadas por (Heagerty y Zheng, 2005):

$$\text{sensitivity}^{\mathbb{I}}(c,t) = \Pr(M_{\text{i}} > c | T_{\text{i}} = t)$$
$$\text{specifity}^{\mathbb{D}}(c,t) = \Pr(M_{\text{i}} \leq c | T_{\text{i}} > t)$$

Notar que la interpretación de las métricas es la usual al considerar la relación entre el predictor (en este caso el risk score) y el punto de corte c. Sin embargo, debido a la naturaleza de los datos de supervivencia[2], se observan cambios en el valor condicional: se presencia un *caso* cuando el cliente incumple en el tiempo t; y un *control*, cuando el cliente aún no ha fallado en t.

Lo anterior responde a un enfoque intensivo-dinámico para la definición de casos y controles, en el cual sólo interesa verificar la fracción de la población (en riesgo) que incumple en un punto en el tiempo (descartando aquellos que ya incumplieron en el pasado) y la parte que lo hará desde ese punto en adelante (enfoque intensivo). Además, a medida que pasa el tiempo, variará tanto la población que incumple como aquella que no (enfoque dinámico).[3].

[2]Los datos de supervivencia incorporan información de tiempo y de estatus, en cambio, tipos de datos más simples sólo entregan información del estatus.

[3]Existen otros enfoques para determinar ROCs en el tiempo, tales como: el acumulativo-dinámico y el

El enfoque intensivo-dinámico es recomendable para evaluar modelos de supervivencia que se utilizarán como insumo para decisiones comerciales preventivas sobre clientes en riesgo de incumplir en un futuro próximo, donde no es de interés considerar a aquellos clientes que ya han incumplido.

Además, este enfoque recrea la misma lógica planteada en la estimación de los modelos de Cox (la cual se basa en los conjuntos de riesgo)[4] y sienta las bases teóricas para lidiar con predictores/covariables dependientes del tiempo (Saha y Heagerty, 2010).

Sin embargo, si se considera una cantidad importante de tiempos de falla (89 en el caso de la memoria), no es factible analizar las curvas ROC, por los que es necesario tener una medida global de discriminación de los modelos, la cual contemple toda la historia. De esta forma, Heagerty y Zheng (2005) propusieron una nueva medida de concordancia basada en el promedio ponderado de las áreas bajo las ROC (los AUC) en el tiempo:

$$\text{iAUC} = C^{\tau} = \int_0^{\tau} \text{AUC}(t) \cdot \omega^{\tau}(t) \mathrm{d}t \tag{12.13}$$

Donde $\omega^{\tau}(t) = 2f(t) \cdot S(t)/W^{\tau}$ y $W^{\tau} = \int_0^{\tau} 2f(t) \cdot S(t)/W^{\tau} \mathrm{d}t = 1 - S^2(t)$.

Esta es una medida de concordancia pues nace de computar la probabilidad $\Pr(M_j > M_k | T_j < T_k, T_j < \tau)$. En la práctica, presenta diferencias (pequeñas) con el C-index de Harrell, dado que establece un factor de corrección/reescalamiento W^{τ} sobre los pesos de los AUC, ello para permitir el uso de horizontes de evaluación finitos (i.e. evaluar todos los tiempos de supervivencia entre 0 a τ años).

intensivo-estático. En el primero se define como un caso a aquella persona que incumple hasta un tiempo t y un control cuando lo hace después. En el segundo, la persona que incumple en t es un caso, mientras que un control corresponde a una persona que incumple en un t^* muy lejano, regla que no cambia incluso al variar t. Del enfoque acumulativo-dinámico cabe destacar que su uso es recomendable para computar la curva ROC para un tiempo t' específico.

[4]Recordar que la estimación requiere emplear la verosimilitud parcial, la cual se basa en la comparación de los risk scores de todos los individuos que están en riesgo a lo largo del tiempo.

Bibliografía

Andreeva, G. (2006). European generic scoring models using survival analysis. *The Journal of the Operational Research Society*, *57*(10), 1180-1187.

Banasik, J., Crook, J., y Thomas, L. (1999). Not if but when will borrowers default. *The Journal of the Operational Research Society*, *50*(12), 1185-1190.

Basel Committee On Banking Supervision. (2000). *Principles for the management of credit risk.*

Basel Committee On Banking Supervision. (2005). *An explanatory note on the basel II irb risk weight functions.* CH-4002 Basel, Switzerland: Bank for International Settlements Press and Communications.

Basel Committee On Banking Supervision. (2015). *Guidance on accounting for expected credit losses.*

Bellotti, T., y Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society*, *60*(12), 1699-1707.

Bogren, F. (2015). *Estimating the term structure of default probabilities for heterogeneous credit portfolios* (Tesis de Master no publicada). Royal Institute of Technology, Suecia.

Bravo, C., Thomas, L. C., y Weber, R. (2012). *Survival analysis for credit scoring with multiple types of defaulters.*

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89–99.

Collett, D. (2015). *Modelling survival data in medical research.* CRC press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, *34*(2), 187-220.

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Grambsch, P. M., y Therneau, T. M. (1994). Proportional hazards test and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515-526.

Harrell, F. E., Lee, K. L., y Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, *15*, 361–387.

Heagerty, P. J., y Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, *61*(1), 92–105.

Hlavac, M. (2015). stargazer: Well-formatted regression and summary statistics tables [Manual de software informático]. Cambridge, USA. Descargado de `http://CRAN.R-project.org/package=stargazer` (R package version 5.2)

IFRS. (2014). *IFRS9 financial instruments project summary.*

Kalbfleisch, J. D., y Prentice, R. L. (2002). *The statistical analysis of failure time data* (2.ª ed.). Wiley.

Kaplan, E. L., y Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457-481.

Klein, J. P., y Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Man, R. (2014). *Survival analysis in credit scoring: A framework for pd estimation* (Tesis de Master no publicada). University of Twente, Holanda.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data : with applications in r*. CRC Press.

Rodríguez, G. (2001). *Lecture notes in parametric survival models.* Princeton University.

Rodríguez, G. (2016). *Lecture notes in generalized linear statistical models.* Princeton University.

Saha, P., y Heagerty, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, *66*(4), 999–1011.

Siddiqi, N. (2006). *Credit risk scorecards: Developing and implementing intelligent credit scoring*. Wiley.

Stepanova, M., y Thomas, L. (2001). Phab scores: proportional hazards analysis behavioural scores. *The Journal of the Operational Research Society*, *52*(9), 1007-1016.

Therneau, T., Crowson, C., y Atkinson, E. (2016, Junio). *Using time dependent covariates and time dependent coefficients in the cox model.* Descargado de `https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf`

Therneau, T. M. (2015). A package for survival analysis in s [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=survival` (version 2.38)

Therneau, T. M., y Grambsch, P. M. (2000). *Modeling survival data: extending the cox model*. Springer Science & Business Media.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149-172.

Tong, E. N., Mues, C., y Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, *218*(1), 132–139.

Xu, R. (2001). *Lecture notes in survival analysis and biostatistical models.* University of California San Diego.