Universidad Nacional Autónoma de México

Facultad de Ciencias

Uso de modelos de supervivencia en la calificación crediticia

T E S I S

Que para obtener el título de
Actuario

Presenta
Jesús Alberto Rodríguez Sánchez.

Director de tesis:
Act. Jaime Vázquez Alamilla.

Ciudad Universitaria, CDMX

2016

A mi madre Rosa María, a mi padre Rodolfo
a mis hermanos Rodolfo, Antonio y Lucero

Agradecimientos

Finalmente, he llegado al desenlace de la primera gran carrera de mi vida. La conclusión de mis estudios, un logro que me trae una dicha enorme y me hace muy feliz, por supuesto todo reconocimiento debe ser compartido con aquellas personas que estuvieron cerca de mi durante este trayecto.

Quiero ser breve, así que empezaré por mencionar a mi hermano Rodolfo que siempre me ha impulsado y me ha brindado esa visión, motivación y energía que cualquier persona necesita para continuar hasta el final.

También agradezco a mi madre que con todo su amor estuvo ahí en las buenas y las malas, en las noches y en las madrugadas, no hay palabras para todo lo que hizo...

Tengo claro todo el apoyo, amor y cariño de mi padre Rodolfo, y mis hermanos Antonio y Lucero. Mis sobrinos también han sido una gran motivación para mi, porque uno siempre quiere ser ejemplo y amigo a la vez... Diego, Frida, Liz, Rosy y Tavo...

A lo largo de mi vida académica conocí compañeros y amigos entrañables: Humberto, Chan, Víctor y Ana, con los que trabajé, estudié, nos desvelamos y pasamos fines de semana compartiendo el mismo objetivo,

el de ser los mejores actuarios. Debo decir que aunque los conocí más tarde, Orlando, Mónica, Karen, Benjamín, Carlos y Lalo también fueron una parte importante de este viaje.

Quiero mencionar a Jaime Vázquez, porque además de ser un gran profesor se convirtió en un gran amigo. Y a Margarita Chávez por todo su apoyo y conocimiento.

Hay muchas personas que estuvieron conmigo en esta carrera, y agradezco por haberlas conocido simplemente porque fueron una parte más de mi y mi carrera: Ana Olmos, Daniel Alfaro, Cristina Collado, Alexander Vargas y Marisol Sánchez.

Sin más, gracias a todos...

Jesús Alberto Rodríguez Sánchez

Índice general

Introducción

El modelo de regresión logística es por excelencia, el modelo estadístico más usado dentro de las instituciones bancarias y otras compañías dedicadas al crédito.

Por otra parte, existe un modelo muy usado dentro del análisis de supervivencia conocido como el modelo de riesgos proporcionales de Cox, que, al igual que todos los modelos tipo regresión, permite establecer un esquema causa-efecto entre una variable de respuesta y un conjunto de variables explicativas. En un estudio de deuda el modelo de Cox permite introducir datos censurados y con ello considerar el tiempo en que tardará en ocurrir un evento, que es exactamente el que se define como el evento de interés, por ejemplo, se puede considerar como el evento de interés el tiempo al default, es decir el momento en que un cliente deja de pagar un crédito, y como covariables la edad, el salario y el número de dependientes, sólo por mencionar algunas. Así, este modelo permitirá hacer inferencias y analizar qué variables serán significativas para modificar estos tiempos que son importantes para que se realice el pago de una deuda.

Para el caso de poblaciones sujetas a un préstamo, en el modelo de riesgos proporcionales de Cox se puede obtener una medición en términos de la posible ganancia máxima que una entidad pueda obtener si elige cierto tipo de características en un cliente que ayudarán a predecir el tiempo en que tardará en dejar de pagar. Este enfoque ha sido utiliza-

do por Stepanova y Thomas (2002); Watkins, Vasnev y Gerlach (2009) y Andreeva (2006), entre otros autores.

En el primer capítulo de este trabajo, se presenta el concepto de calificación crediticia (credit scoring), aspectos generales y consideraciones de negocio, así como la construcción estadística de una scorecard [1]. Se finaliza con la introducción al estudio de las curvas de ROC para la medición del desempeño del modelo.

En el capítulo dos, se hace una revisión de la teoría del análisis estadístico de supervivencia, haciendo énfasis principalmente en el modelo de riesgos proporcionales de Cox, desde su construcción hasta la validación del mismo.

Finalmente, el capítulo tres, da un panorama general de cómo debería desarrollarse un credit scoring y sus respectivas scorecards a través del enfoque de supervivencia, tomando en cuenta la variable default. Se hace una revisión descriptiva de las variables y se da el bosquejo de la construcción de una scorecard, proponiendo un modelo construido a partir del análisis de las variables, dicha scorecard tiene como objeto la toma de decisiones, detallando lo que considera una institución bancaria o de préstamos para este efecto.

En esta tesis se utiliza el software estadístico R así como su paquetería OISurv, que se ha utilizado para el análisis de datos, y gráficas, además de la regresión y sus respectivos residuos.

[1] Una *scorecard* es una tarjeta que se utiliza para escribir las características de un cliente, y con ella se evalúa si el cliente puede ser sujeto a un préstamo o no.

Capítulo 1

Calificación crediticia

1.1. Concepto de calificación crediticia (*credit scoring*)

A lo largo de la historia, la práctica del crédito, ha sido imprescindible y ha evolucionado de manera diferente de acuerdo a la época y al tipo de transacción. En general, un crédito consiste en que el prestamista otorga un monto fijo al prestatario, quien a su vez liquida la deuda mediante uno o más pagos. Uno de los retos de las instituciones financieras es decidir a quién se le debe otorgar un crédito, debido al riesgo que representa el incumplimiento de el o los pagos. Otro punto importante es la cantidad del préstamo que se debe hacer, así como las estrategias que se deben llevar a cabo para que resulte satisfactorio y sea un beneficio para ambas partes.

Tradicionalmente las decisiones para aprobar un crédito a un prestatario se tomaban de acuerdo con una entrevista, o se hacían de una manera muy lenta. Por ello, se ha desarrollado la técnica estadística conocida como calificación crediticia (credit scoring) que se utiliza para decidir si se debe conceder crédito a un solicitante o no.

La manera de operar los sistemas de calificación crediticia es evaluar el riesgo de hacer un préstamo a un prestatario. En caso de que dicha evaluación sea satisfactoria, el resultado culminará determinando que un solicitante es solvente, lo cual significa que tiene un nivel de riesgo bajo de dejar de pagar o de dejar pagos incompletos al prestamista.

Existen diversas técnicas para realizar la calificación crediticia y la mayoría utiliza una muestra grande de clientes con los detalles de su solicitud y su historial crediticio. Con ellas se determinan las conexiones entre las características de los clientes y qué tan buena o mala es su historia subsecuente. Tradicionalmente, esta evaluación se lleva a cabo estimando la probabilidad de que un solicitante deje de pagar o caiga en una condición de ausencia de pagos, que en inglés se conoce como default; en este trabajo, se considerará también el concepto de pago temprano, el cual consiste en liquidar la deuda antes del tiempo establecido; el interés por esta variable tiene que ver con las consecuencias que le ocasionan a un negocio el que un cliente deje de pagar intereses.

Se genera una scorecard que es una tarjeta donde se va obteniendo un puntaje de acuerdo a los atributos de los solicitantes para obtener una calificación final, la cual determina si el riesgo de un cliente es lo suficientemente grande como para otorgarle un préstamo o no. De esta manera se termina eligiendo a los clientes de mayor solvencia, lo cual ha significado un cambio muy importante durante los últimos años en la forma de determinar cuándo caerá en default un cliente. Es posible que si el tiempo al default es lo suficientemente grande, los intereses cobrados durante ese tiempo compensen, o incluso excedan, las pérdidas que resulten del default de toda la cartera de clientes. Otros factores que pueden afectar los beneficios para el prestamista durante una transacción de crédito son: cuando un cliente cierra su cuenta, si liquida su deuda de manera prematura o si cambia de prestamista, por lo que cuando el pago anticipado ocurra, el prestamista perderá una proporción del interés de la deuda.

1.2. Construcción estadística de un Credit Scoring

Los métodos para llevar a cabo la *calificación crediticia* surgieron alrededor de hace 50 años, y eran muy distintos a los que se utilizan actualmente. Los más importantes han conservado su esencia hasta el momento y estos son los *estadísticos*, que se utilizan para desarrollar *scorecards* crediticias. La ventaja es que permiten usar el conocimiento de las propiedades de los estimadores muestrales así como de las herramientas de los intervalos de confianza y de las pruebas de hipótesis en el contexto de la calificación crediticia. El uso de la estadística es de gran utilidad, pues se pueden identificar y remover ciertas características dejando paso para trabajar únicamente con las importantes dentro de una scorecard. De esta manera, las preguntas que los futuros clientes responden en alguna entrevista pueden ser modificadas consiguiendo mejores características.

Una primera opción para decidir cuáles características son relevantes es buscar la regla que minimiza el costo esperado de decidir si se acepta a un nuevo cliente o no, mientras que una segunda opción es el enfoque de Fisher, el cual se puede ver como una forma de regresión lineal, y esto conlleva a una investigación de otras formas de regresión que tienen supuestos menos restrictivos que garantizan su optimalidad y desempeño, además conducen a las reglas de puntaje lineal. Por mucho, la más exitosa es la regresión logística, siendo el método estadístico más común para la calificación crediticia.

El método utilizado en este trabajo, es el criterio de clasificación de Stepanova y Thomas (2002), donde se utiliza el enfoque de supervivencia para revisar las características de los clientes y con ellas poder obtener partes específicas que sean determinantes para la calificación de dicho cliente.

1.2.1. Diseño y uso de una *Scorecard* (tarjeta de puntaje)

Una scorecard o tarjeta de puntaje, es una herramienta que se utiliza para evaluar el nivel de riesgo asociado a un cliente o un solicitante de un préstamo, y lo hace a través de la medición estadística de probabilidades o puntajes que están basadas en aspectos tales como tasas de aprobación de productos, beneficios y pérdidas, y con ellas tener una base para tomar decisiones. Se deben determinar estadísticamente diversas características o grupos de ellas que tengan poder de discriminación para separar a los clientes considerados como buenos y malos.

La elaboración de la scorecard es una tarea que, además de ser costosa, lleva una serie de implicaciones que involucran a todas las áreas de una institución de crédito, de tal forma que el proceso de construcción y definición de todas las aristas que ayudan a tomar la decisión de aceptar o rechazar al solicitante, cumpla con todos los objetivos que la institución necesite y que se lleve a cabo de tal forma que simule una evaluación personal que será completamente insesgada.

Asimismo para poder decidir y hacer una buena scorecard se deben tomar en cuenta muchos aspectos como la probabilidad de *default*, una definición de mal comportamiento y periodos distintos de tiempo donde el cliente incumple con sus pagos. Por supuesto todo dependerá también del tipo de préstamo que se esté considerando, ya sea una hipóteca, una tarjeta de crédito, un préstamo personal u otro.

Finalmente, el uso de las scorecards y en general del credit scoring tienen un potencial muy grande puesto que puede ser muy extensivo; brinda diversas opciones para que el prestatario pueda detectar (antes, durante y después del ciclo del crédito) a los clientes buenos y malos, evaluarlos y tomar decisiones con respecto a su tratamiento. Es decir, evaluando y revisando el comportamiento de los clientes, sus niveles de morosidad, rentabilidad y dependiendo de los diversos comportamientos, se llevan

a cabo acciones para favorecer a los clientes y al mismo tiempo al prestatario, al ofrecer descuentos, promociones, reducciones o incrementos de líneas crediticias, nuevos productos o algún reajuste de su deuda para que pueda liquidarse toda y con ello optimizar ganancias y reducir riesgos de incumplimiento.

1.2.2. Clasificación ordinaria de las características

Además de la descripción de los métodos y modelos con los que se puede construir una *scorecard*, es necesario también ver lo que ocurre con los datos. Un aspecto importante a considerar es cómo clasificar las características, para lo que es necesario determinar subpoblaciones. En general existen muchas situaciones en las que cada elemento de un conjunto de objetos pertenece a dos o más clases, y en un estudio estadístico se quiere obtener una buena separación, por lo que se puede realizar algún procedimiento de asignación basado en la información observada acerca del objeto.

Algunos ejemplos prácticos:

i) La conducción de un diagnóstico médico, en el cual el objetivo es diagnosticar a cada paciente la enfermedad A o la enfermedad B.

ii) Desarrollar sistemas de reconocimiento de discursos, en el cual el objetivo sea clasificar palabras pronunciadas.

iii) Evaluar aplicaciones financieras crediticias, en las cuales el objetivo es asignar a cada solicitante ya sea en la clase de tener un probable *default* o un *improbable default*.

iv) Filtrar mensajes de correo electrónico, para decidir si son "spam" o son mensajes genuinos.

v) Examinar transacciones de tarjetas de crédito, para decidir si son o no fraudulentas.

Por supuesto, los que competen a este trabajo tienen que ver con las características de ser o no un buen sujeto de crédito, para poder otorgar un préstamo, por lo que a continuación se describen algunos de los métodos más comunes.

1.2.2.1. Determinación de subpoblaciones

Un uso importante de las características es que se pueden separar las poblaciones en subpoblaciones, y se pueden construir diferentes *scorecards* de cada una de estas supoblaciones, y las razones pueden ser por políticas definidas por parte del prestatario, o bien, razones de corte estadístico.

Las razones estadísticas de separar la población en subpoblaciones se debe a la gran cantidad de interacciones entre las características, además de la sensibilidad de construir una scorecard para cada uno de los diferentes atributos de esa característica. Por otro lado, podría suceder que cuando se segmenta una población no necesariamente se garantiza una mejora en la predicción. Si las subpoblaciones no son muy distintas, no es conveniente redefinirlas, pues con muestras más pequeñas se podría perder el desempeño del modelo que se considere.

1.2.2.2. Clasificación ordinaria

Una vez determinadas las subpoblaciones, es necesario que para cada característica se puedan dividir las posibles respuestas en un número pequeño de clases, esto es, hacer una *clasificación ordinaria*. El uso de esta clasificación es indispensable puesto que se pueden tener variables categóricas (donde se tiene un conjunto de respuestas discretas) o variables continuas (donde el conjunto de respuestas puede ser infinito).

1.2.2.3. Estadística Ji-cuadrada

Sean g_i y b_i los números de sujetos *buenos* y *malos* con el atributo i, y sean g y b el número total de buenos y malos. Entonces,

$$G = \frac{(g_i + b_i)g}{g+b} \qquad \text{y} \qquad B = \frac{(g_i + b_i)b}{g+b}$$

son el número esperado de buenos y malos con el atributo i, es decir, el cociente para este atributo es el mismo que para la población completa. Entonces

$$s^2 = \sum_i \left(\frac{(g_i - G)^2}{G} + \frac{(b_i - B)^2}{B} \right)$$

que es la estadística con distribución χ^2. Formalmente, mide qué tan probable es que no exista diferencia en el cociente *bueno:malo* en las diferentes clases, y puede compararse con la distribución χ^2 con $k-1$ grados de libertad, donde k es el número de clases de las características. Sin embargo, se puede usar como una medida para ver qué tan diferentes son las ventajas en las diferentes clases con un valor más alto, reflejando grandes diferencias en éstas.

1.2.2.4. Estadística de información

La *estadística de información* está relacionada con las medidas de entropía que aparecen en la teoría de la información y se define como:

$$I = \sum_{\forall i} \left(\frac{g_i}{g} - \frac{b_i}{b} \right) \log \left(\frac{g_i b}{b_i g} \right)$$

En estadística, también se conoce como divergencia o valor de información y busca identificar qué tan diferentes son $\mathbb{P}(x|g)$ y $\mathbb{P}(x|b)$ cuando x toma valores del atributo i. La motivación de la estadística de información es de la forma:

$$\sum_i \left(\frac{g_i}{g}\right) \log\left(\frac{g_i}{g}\right)$$

cuando hay g observaciones y g_i de ellas que son del tipo i es como sigue. El número de formas en las que esta distribución ocurre es $N_g = \frac{g!}{g_1!g_2!\ldots g_p!}$ si hay p tipos en total. La información se obtiene como el log del número de formas distintas en que se puede obtener el mensaje que fue visto, de tal forma que $I_g = \log(N_g) = \log(g!) - \sum_i \log(g_i!) \approx g\log(g) - \sum_i g_i \log(g_i)$. La información promedio es

$$\frac{I_g}{g} \approx -\sum_i \left(\frac{g_i}{g}\right)(\log(g_i) - \log(g)) = -\sum_i \left(\frac{g_i}{g}\right)\log\left(\frac{g_i}{g}\right)$$

Esta estadística de información se puede pensar como la diferencia de la información en los buenos y la información en los malos, es decir,

$$-\sum_{\forall i}\left(\frac{g_i}{g} - \frac{b_i}{b}\right)\left(\log\left(\frac{g_i}{g}\right) - \log\left(\frac{b_i}{b}\right)\right)$$

Cuando hay grandes diferencias entre $\mathbb{P}(x|g)$ y $\mathbb{P}(x|b)$ producen valores grandes de I y corresponden a características que son más útiles para diferenciar a los buenos de los malos.

1.2.2.5. Estadística D de concordancia de Somer

La prueba de concordancia de la estadística D de Somer supone que las clases de las características están ordenadas de menor a mayor, la cual, tiene la menor tasa de *buenos*, a la mayor, la cual tiene la tasa más alta de *buenos*. La estadística de concordancia describe si se elige uno bueno aleatoriamente del conjunto de *buenos* y un malo aleatoriamente del conjunto de *malos*, el atributo del malo, x_b, estará en una clase menor que el atributo del bueno, x_g. Mientras más alta es esta probabilidad, hay un mejor orden de las clases de las características que refleja la división buenos-malos en la población. La definición precisa de D, es

el *pago* esperado de una variable, el cual es 1 si el orden pone al malo debajo del bueno, y -1 si el orden pone al bueno debajo del malo, y 0 si son iguales:

$$D = 1 \cdot \mathbb{P}(X_b < X_g) - 1 \cdot \mathbb{P}(X_b > X_g) + 0 \cdot \mathbb{P}(X_b = X_g)$$

$$= \sum_i \frac{\left(\sum_{j<i} b_j\right) g_i - \left(\sum_{j<i} g_j\right) b_i}{bg}$$

1.2.2.6. Clasificación ordinaria usando el enfoque de análisis de supervivencia

Se utilizará el criterio de clasificación de Stepanova y Thomas (2002), donde para asegurarse que los sistemas de *calificación crediticia* predicen más acerca de los datos de lo que estos sistemas los describen, las características continuas tales como la edad, se dividen en *clases*, y los valores de las características discretas con muchos valores se agrupan del mismo modo, y cada *clase* se reemplaza con una variable muda o *dummy* binaria.

Los enfoques tradicionales para encontrar las divisiones más apropiadas involucran revisar el cociente *bueno:malo* o medidas relacionadas para diferentes valores de la característica, y entonces, agrupando los valores con cocientes de *bueno:malo* similares. La elección de un horizonte de tiempo es inherente en estos enfoques, de tal forma que los *defaults* antes de ese horizonte de tiempo son malos, mientras que los *defaults* después de éste, son buenos.

En el contexto de supervivencia es más apropiado usar un enfoque que evite la necesidad de usar un horizonte de tiempo arbitrario. También es el caso en el que los modelos de supervivencia se construyen para

estimar el riesgo, donde se utilizan las características de *default* y *pago temprano*, entonces se deben hacer los rangos de las variables de manera distinta para los diferentes riesgos.

Por las razones anteriores es mucho más apropiado utilizar el enfoque de supervivencia para la *clasificación ordinaria* de las características.

El siguiente método se puede utilizar para características continuas:

1. Dividir las características entre 5 y 10 rangos iguales (dependiendo del estudio).
2. Definir una variable binaria para cada rango.
3. Ajustar el modelo de riesgos proporcionales de Cox con estas variables binarias.
4. Trazar los estimadores de los parámetros para todos los rangos.
5. Elegir los rangos basados en la similitud de los estimadores de los parámetros.

1.2.3. Medición del desempeño del modelo

La evaluación del desempeño del modelo de la construcción de una *scorecard* es imprescindible, pues surge la pregunta natural de si el modelo es bueno, o qué tan bueno es. Existen diferentes maneras de medir el desempeño del modelo. Para hacerlo se debe utilizar la muestra con la que se construyó el modelo, y algunas veces se hace la prueba en otra muestra para revisar que efectivamente el desempeñõ es similar, esta última muestra se conoce como la *muestra de retención.*

Cuando la cantidad de datos es limitada, es decir, si se tuviera, por ejemplo, una cartera nueva de clientes, se puede probar la validez del sistema por métodos que construyen y prueban esencialmente el mismo

conjunto de datos, pero sin causar errores de clasificación para que fuese sesgado; para ello sirven los métodos de *validación cruzada*, *leave one out*, *bootstrap* y *jackknife*.

Otro tipo de recurso que se necesita es la forma en que se pueden describir dos poblaciones distintas y sus características, se puede medir qué tan bien se separan los dos grupos, utilizando la *distancia de Mahalanobis* y la estadística de *Kolmogorov Smirnov (K-S)*. Sin embargo, se puede extender el enfoque de *K-S* para obtener una buena separación de los dos grupos sobre el rango completo de las calificaciones. Una forma de hacerlo es generando una curva, conocida como la *curva de la característica receptora de operación* o curva *ROC* (Receiver Operating Characteristic), que es muy útil para comparar las propiedades de clasificación de las diferentes *scorecards*.

1.2.3.1. Curvas ROC

Una vez que se han clasificado las características, es necesario evaluar la calidad del desempeño del procedimiento de clasificación y revisar si el desempeño es lo suficientemente bueno, de lo contrario mejorarlo o incluso reemplazarlo.

El enfoque que se describe en este trabajo para hacer una evaluación a estos procedimientos y que además es ampliamente usado en la práctica, es el de la *curva ROC*. Idealmente, se debería elegir el criterio particular del desempeño que corresponde a un aspecto que se considera central de algún método específico. Sin embargo, la mayoría de las veces es difícil identificar un sólo aspecto central, y en cualquier caso pueden no conocerse las circunstancias precisas bajo las cuales la regla se aplicará en el futuro, pues es posible que sufra cambios. Para tal efecto, es muy útil tener una manera de mostrar y resumir el desempeño sobre un amplio rango de condiciones, que es exactamente el que provee la *curva ROC*.

1.2.3.2. Definición de las curvas de ROC

Supongamos que se tiene una población la cual se va a separar en dos subpoblaciones a través de una regla de clasificación, la manera para hacerlo es a través de un umbral T, entonces si se tiene un valor t de este umbral, para cada individuo cuyo puntaje s excede a t será colocado en la población digamos, P o en la población N si ocurre lo contrario. Para poder evaluar la eficacia de esta regla de clasificación se necesita calcular la probabilidad de hacer una colocación incorrecta. Con ello, se define lo siguiente:

1. La probabilidad de que un individuo de la población, P, se coloque correctamente, es decir, la tasa de verdaderos positivos es: $tp = \mathbb{P}(s > t|P)$.

2. La probabilidad de que un individuo de la población, N, se coloque incorrectamente, es decir, la tasa de falsos positivos $fp = \mathbb{P}(s > t|N)$.

3. La probabilidad de que un individuo de la población, N, sea correctamente clasificado, es decir, la tasa de verdaderos negativos: $tn = \mathbb{P}(s \leq t|N)$.

4. La probabilidad de que un individuo de una población, P, sea incorrectamente clasificado, es decir, la tasa de falsos negativos: $fn = \mathbb{P}(s \leq t|P)$.

La descripción completa acerca del desempeño de la regla de clasificación se completa con las cuatro probabilidades anteriores y el valor del umbral t. Es claro que para un buen desempeño se requieres altas tasas de verdaderos y bajas tasas de falsos. Sin embargo, esto es para una elección de un umbral o un punto de partida t, cuya elección generalmente no es conocida, pero debe ser determinada como parte de la construcción del clasificador. Variaciones de t y la evaluación de las cuatro probabilidades antes descritas dan toda la información necesaria,

pero como $tp + fn = 1$ y $fp + tn = 1$ no es necesario utilizar las cuatro probabilidades, basta con graficar la curva que se obtiene al variar t, pero utilizando únicamente las tasas de verdaderos positivos y verdaderos negativos (fp, tp) como puntos, contra ejes ortogonales donde fp es el valor de las abscisas y tp el valor de las ordenadas.

1.2.3.3. Características generales

La curva ROC es una curva continua que se ubica en el cuadrado unitario que es el que está definido por los cuatro puntos: $(0,0)$, $(1,0)$, $(1,1)$ y $(0,1)$, y esta curva tiene dos posibilidades extremas y poco comunes, a partir de las cuales se puede entender la evaluación del desempeño del modelo: una donde $\mathbb{P}(s|P) = \mathbb{P}(s|N) = \mathbb{P}(s)$ donde ocurre que las poblaciones son exactamente las mismas, y la curva ROC asociada es una línea recta que va desde el punto $(0,0)$, al $(1,1)$, y usualmente se conoce como *diagonal de oportunidad*, puesto que representa esencialmente una localización aleatoria de individuos de las dos poblaciones. Y en el otro extremo (que es usualmente imposible) hay una separación completa de la $\mathbb{P}(s|P)$ y la $\mathbb{P}(s|N)$, y es cuando la curva ROC pasa a lo largo de los bordes superiores de la gráfica: una línea recta de $(0,0)$ a $(0,1)$ seguida por una línea recta de $(0,1)$ a $(1,1)$.

De este modo la curva ROC cae sobre el triángulo superior de la gráfica. Mientras esté más cerca de la esquina superior izquierda, más cerca se tiene una situación completa de separación de poblaciones, y con ello un mejor desempeño del modelo clasificador. No hay pérdida de generalidad si se supone que la curva siempre cae en la parte superior del triángulo.

1.2.3.4. Puntajes continuos

Todo lo que se ha desarrollado acerca de las curvas de ROC puede ser aplicado a cualquier tipo de escala de medición para los puntajes S. Sin

embargo, para un enfoque continuo se hace una forma más conveniente de la curva ROC. Como la curva ROC es la gráfica de $y(t) = h(x(t))$ a través de los valores de t, donde $x(t) = \mathbb{P}(s > t|N)$ y $y(t) = \mathbb{P}(s > t|P)$. Cuando S es una variable continua, tendrá funciones de densidad y distribución para cada una de las dos poblaciones. Sean f y F la función de densidad y de distribución respectivamente, para N, y sean g y G las correspondientes para P. Entonces, $\mathbb{P}(s|N) = f(s)$, $\mathbb{P}(s|P) = g(s)$, $x(t) = 1 - F(t)$ y $y(t) = 1 - G(t)$. Por lo tanto, se puede obtener:

$$y = 1 - G(F^{-1}(1 - x)) \qquad \text{para} \qquad 0 \leq x \leq 1 \tag{1.1}$$

la cual será la ecuación más conveniente para la curva ROC con puntajes o calificaciones continuas.

1.2.3.5. Pendiente de la curva ROC y resultados óptimos

Existen dos resultados interesantes, uno proveniente de la "teoría de la decisión" y el otro de la teoría de "pruebas de hipótesis".

Costo esperado de tasas de mala calificación

Los errores de clasificación ocurren cuando se coloca a un individuo que pertenece a N en P, o viceversa, y si el umbral t se usa en los puntajes del clasificador, entonces sus probabilidades respectivas, son $x(t)$ y $1 - y(t)$. Supóngase que los costos de cometer estos errores son $c(P|N)$ y $c(N|P)$, y que las proporciones relativas de los individuos P y N en la población son q y $1 - q$ respectivamente. Entonces, el costo esperado dada una mala clasificación cuando se usa el clasificador en el umbral t sería:

$$C = q\,[1 - y(t)]\,c(N|P) + (1 - q)x(t)c(P|N) \tag{1.2}$$

Puesto que la pendiente de la curva ROC es monótona decreciente, el umbral que minimiza este costo se obtiene resolviendo:

$$\frac{dC}{dx} = 0 \tag{1.3}$$

y una diferenciación sencilla muestra que este umbral corresponde a la pendiente:

$$\frac{dy}{dx} = \frac{(1-q)c(P|N)}{qc(N|P)} \tag{1.4}$$

de la curva ROC. El caso particular en el que los dos costos son iguales, o sea $q = 0.5$, entonces el umbral "óptimo" (es decir, la tasa que proporciona el error mínimo total) estará en el punto donde la curva ROC tiene pendiente 1, es decir, es paralela a la *diagonal de oportunidad*.

Capítulo 2

El análisis estadístico de supervivencia

2.1. Introducción al análisis de supervivencia

El análisis de supervivencia es una área de la estadística de gran importancia y aplicación, es utilizado principalmente por biomédicos, epidemiólogos, ingenieros, consultores estadísticos y también actuarios.

Un modelo de análisis de supervivencia es simplemente una función de densidad de probabilidad de una variable aleatoria del tiempo de supervivencia. Donde el tiempo de supervivencia se puede definir como el tiempo hasta la ocurrencia de un evento previamente definido, este evento se conoce como *falla*. Así pues, en cualquier situación que tenga que ver con un modelo de supervivencia, siempre se tendrá una persona u objeto definido y un concepto de supervivencia asociado y por ello, de *falla* de la persona u objeto en cuestión. Se presentan a continuación algunos ejemplos de *fallas* y sus variables aleatorias asociadas:

1. El tiempo de vida de operación de un foco de luz. Aquí se dice que el foco sobrevive mientras continúe alumbrando y *falla* en el

instante en que se quema o se funde.

2. El tiempo de vida de una persona. La persona sobrevive desde su nacimiento o desde alguna edad preestablecida, hasta que ocurre la muerte, lo cual constituye la falla de la persona.

3. El tiempo hasta que un prestatario deja de pagar una deuda con alguna entidad prestamista; ya sea bancaria o con otra institución que le hizo un préstamo que debía ser liquidado en una serie de pagos en determinado tiempo. La deuda subsiste mientras la persona puede hacer efectivos los pagos, hasta que en determinado momento deje de pagarlos por alguna cuestión personal, o porque ha terminado de pagar la deuda, lo cual constituye la falla del prestatario.

2.1.1. Datos censurados

Un estudio estadístico de supervivencia, es corto por motivos de costo, falta de muestras, cuestiones de tiempo, o simplemente la complicación natural del experimento, lo que puede generar datos u observaciones censuradas.

Si no se observa la *falla* en el periodo de tiempo del estudio, o se observa una falla que no es de interés, se le conoce como tiempo de supervivencia *censurado*, o simplemente tiempo de censura o censura. Cuando las observaciones no son censuradas se llaman observaciones completas o tiempos de *falla*.

Por ejemplo, si se mide el tiempo hasta que ciertos pacientes fallecen o contraen cierta enfermedad, algún paciente puede estar vivo, o sin señales de enfermedad al final del periodo de estudio, lo que hace que los tiempos exactos de *falla* de la supervivencia sean imposibles de conocer. Otro ejemplo clásico en seguros de no vida, es cuando un seguro está topado a cierto límite de póliza, en el momento en el que la aseguradora

toma las muestras de los costos, es decir, de la *severidad* de los siniestros, puede ocurrir que el costo de un siniestro sea mayor al límite de póliza, de tal forma que al contar las severidades, el valor que se tomará en cuenta será justamente el límite de la póliza y no el verdadero valor del siniestro.

Se definirán tres tipos de datos censurados:

Censura tipo I. Se consideran las *fallas* en un estudio desde $t = 0$ hasta el final del estudio t_f. Las fallas se van contabilizando y es posible que algunas observaciones tarden más tiempo del establecido en el estudio, estas observaciones se conocen como censura tipo I y su tiempo de censura es exactamente t_f.

Censura tipo II. Cuando se realiza un estudio de supervivencia, a veces es necesario cubrir un mínimo de fallas para poder describir mejor lo que se está estudiando, por ello en la censura tipo II se establece al inicio del estudio un número específico d de *fallas*, de tal forma que el estudio comienza en $t = 0$ y finaliza en cuanto se hayan observado las d fallas.

Censura aleatoria El estudio comienza en $t = 0$ y termina en $t = t_f$, las observaciones entran al estudio en cualquier momento y no tienen un momento específico de salida. Se contabilizan las fallas hasta t_f y las censuras ya sea que hayan ocurrido antes de finalizar el estudio tomando t_c como su tiempo de censura, o si terminó el estudio y su tiempo de censura es en $t_f = t_c$. Este tipo de censura es frecuente en estudios clínicos donde cada paciente es tratado con una terapia distinta.

2.1.2. Funciones para analizar el tiempo de supervivencia

Sea T el tiempo de supervivencia. La supervivencia de T se caracteriza por:

Definición. La *función de supervivencia* se denota por $S_T(t)$, y se define como la probabilidad de que un individuo sobreviva más allá del tiempo t, es decir,

$$S_T(t) = \mathbb{P}(T > t)$$

Además la función de supervivencia debe satisfacer:

i) Es una función no creciente, es decir, sean $a, b \in \mathbb{R}$, tal que si $a \leq b$ entonces $S(a) \geq S(b)$.

ii) Es continua por la izquierda.

iii)
$$\lim_{t\to 0} S_T(t) = 1 \qquad \text{y} \qquad \lim_{t\to\infty} S_T(t) = 0$$

De esta manera la *función de distribución* de la variable aleatoria T (definida anteriormente), se denota como $F_T(t)$, que es la probabilidad de ser menor a un tiempo dado t, es decir,

$$F_T(t) = \mathbb{P}(T \leq t) = 1 - S_T(t) \tag{2.1}$$

Por su parte, la *función de densidad* de T, $f_T(t)$, si T es continua, es la derivada de la función de distribución de T $F_T(t)$, o equivalentemente el negativo de la derivada de la función de supervivencia. Esto es,

$$f_T(t) = \frac{d}{dt}F_T(t) = -\frac{d}{dt}S_T(t)$$

Definición. La *función de tasa de riesgo*, recibe su nombre puesto que se define como la probabilidad de *falla* durante un intervalo de tiempo muy pequeño, suponiendo que el individuo está vivo al inicio de ese intervalo, o de otro modo:

$$h_T(t) = \lim_{\Delta t\to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f_T(t)}{S_T(t)} \tag{2.2}$$

También se conoce como *fuerza de mortalidad* o *tasa instantánea de falla.*

Definición. La *función de riesgo acumulado* está definida como:

$$H_T(t) = \int_{-\infty}^{t} h_T(x)dx$$

2.1.3. Relaciones importantes entre las funciones de supervivencia

Además de la importancia que tienen las funciones previamente descritas, éstas están relacionadas entre sí. A continuación se presentan las relaciones más importantes entre estas funciones.

1. La función de riesgo, también puede escribirse como:

$$h_T(t) = \frac{f(t)}{S_T(t)} \tag{2.3}$$

2. Ya que la función de densidad es la derivada de la función de distribución, tenemos:

$$f_T(t) = \frac{d}{dt}\left[F_T(t)\right] = \frac{d}{dt}\left[1 - S_T(t)\right] = -\frac{d}{dt}S_T(t) \tag{2.4}$$

3. Sustituyendo (2.4) en (2.3) se sigue:

$$h_T(t) = -\frac{\frac{d}{dt}S_T(t)}{S_T(t)} = -\frac{d}{dt}\log S_T(t)$$

4. Integrando la función de riesgo (2.3) de cero a t y recordando que $S_T(0) = 1$, tenemos:

$$-\int_0^t h_T(x)dx = \log S_T(t)$$

ó

$$H_T(t) = -\log S_T(t)$$

ó

$$S_T(t) = e^{-H_T(t)}$$

$$= \exp\left(-\int_0^t h_T(x)dx\right) \tag{2.5}$$

5. De (2.3) y (2.5) tenemos:

$$f_T(t) = h_T(t) \cdot \exp[-H_T(t)] \tag{2.6}$$

Existen métodos paramétricos de estimación para el ajuste de modelos y para identificar factores significativos de pronóstico. Estos métodos pueden utilizarse si se conoce la distribución de T, sin embargo, la mayoría de las veces esta distribución se desconoce por lo que el uso de las técnicas de las que se habla se ve muy limitado para realizar cualquier tipo de inferencia. Los métodos tanto paramétricos como no paramétricos para estimar la función de supervivencia, pueden encontrarse en libros estándar de análisis de supervivencia (por ejemplo: Collet (2009) y Kleinbaum (2012)). El estimador no paramétrico de $S_T(t)$ más conocido es el estimador de Kaplan-Meier. El modelo de regresión más conocido dentro del análisis de supervivencia es el modelo de riesgos proporcionales de Cox, en el cual no es necesario especificar la distribución de T y con ello se vuelve un modelo semiparamétrico, en donde se estiman los coeficientes de regresión.

2.2. Estimador de Kaplan Meier

El estimador paramétrico más conocido de la función de supervivencia, es el estimador *producto límite*, mejor conocido como el *estimador de*

Kaplan-Meier (K-M), el cual ajusta la función de supervivencia empírica, de tal forma que refleja la presencia de las observaciones censuradas.

Recordando la definición de censura por la derecha. Sean T_i y C_i los tiempos de falla y de censura de cada individuo $i = 1, \ldots, n$, con funciones de densidad y distribución definidas para cada i. Se observa el par (Y_i, δ_i) donde

$$Y_i = min(T_i, C_i) \qquad y \qquad \delta_i = \begin{cases} 1 & \text{si } T_i \leq C_i \\ 0 & \text{si } C_i < T_i \end{cases}$$

de tal forma que $y_{(i)}$ denota la i-ésima observación ya sea censurada o no. Asimismo se denotará

$\mathbf{R}(y_{(i)})$ al conjunto de individuos en riesgo, es decir, todas las observaciones que continúan en el estudio sin haber presentado falla o censura, justo antes del tiempo t.

n_i será el número de individuos en $\mathbf{R}(y_{(i)})$ (ya sean tiempos de falla o de censura).

d_i el número de fallas al tiempo $y_{(i)}$

p_i es $\mathbb{P}(T > y_{(i)}|T > y_{(i-1)})$, es decir, la probabilidad de que un individuo sobreviva en el intervalo (t_i, t_{i-1}) dado que estaba vivo al comienzo del intervalo.

$q_i = 1 - p_i$

Recordando la regla general de la multiplicación para eventos conjuntos A_1 y A_2

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_2|A_1)\mathbb{P}(A_1)$$

Aplicando repetidamente esta regla para la función de supervivencia, se expresa

$$S(t) = \mathbb{P}(T > t) = \prod_{y_{(i)} \leq t} p_i$$

y los estimadores de p_i y q_i son

$$\hat{q}_i = \frac{d_i}{n_i} \qquad y \qquad \hat{p}_i = 1 - \hat{q}_i = 1 - \frac{d_i}{n_i} = \left(\frac{n_i - d_i}{n_i}\right)$$

Por lo tanto, el estimador de *Kaplan-Meier* de la función de supervivencia es

$$\widehat{S}(t) = \prod_{y_{(i)} \leq t} \widehat{p}_i = \prod_{y_{(i)} \leq t} \left(\frac{n_i - d_i}{n_i}\right) = \prod_{i=1}^{k} \left(\frac{n_i - d_i}{n_i}\right) \tag{2.7}$$

donde $y_{(k)} \leq t < y_{(k+1)}$.

Además de estos estimadores, también se tienen estimadores para la varianza de la función de supervivencia, y con ello, se pueden obtener los intervalos de confianza respectivos. A continuación se presenta la fórmula de Greenwood (1926) para la varianza del estimador de K-M

$$var(\widehat{S}(t)) = \widehat{S}^2(t) \sum_{y_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)} = \widehat{S}^2(t) \sum_{i=1}^{k} \frac{d_i}{n_i(n_i - d_i)} \tag{2.8}$$

donde $y_{(k)} \leq t < y_{(k+1)}$.

Además, para cada t fijo,

$$\widehat{S}(t) \sim N\left(S(t), \widehat{var}(\widehat{S}(t))\right)$$

por lo tanto, al tiempo t, un intervalo aproximado del $100 \times (1 - \alpha)\,\%$ de confianza para la probabilidad de supervivencia $S(t) = \mathbb{P}(T > t)$ está dado por

$$\widehat{S}(t) \pm z_{\alpha/2} \times s.e.\left(\widehat{S}(t)\right) \tag{2.9}$$

donde $s.e.\left(\widehat{S}(t)\right)$ es la raíz cuadrada del estimador de la varianza de la fórmula de Greenwood.

2.2.1. Estimadores del riesgo

Es importante considerar estimadores no paramétricos de la función de riesgo y riesgo acumulado, por ello se presentan los más utilizados; considérese t_i un tiempo de falla ordenado y sin repetición tal que $i = 1, 2, \ldots, k \leq n$, con ello se tiene para $h(t)$

- El estimador de la función de riesgo en el intervalo $t_i \leq t < t_{i+1}$

$$\hat{h}(t) = \frac{d_i}{n_i(t_{i+1} - t_i)}$$

  Este se conoce como el estimador del tipo *K-M*. Estima la tasa de muerte por unidad de tiempo en el intervalo $[t_i, t_{i+1})$.

- La estimación de la función de riesgo en un tiempo de falla observado t_i

$$\tilde{h}(t_i) = \frac{d_i}{n_i} \tag{2.10}$$

y para la función de riesgo acumulado $H(t)$:

- El estimador construido con *K-M*

$$\widehat{H}(t) = -\log \widehat{S}(t) = -\log \prod_{y_{(i)} \leq t} \left(\frac{n_i - d_i}{n_i} \right)$$

$$\widehat{var}(\widehat{H}(t)) = \sum_{y_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

- El estimador de *Nelson-Aalen* (1972, 1978)

$$\tilde{H}(t) = \sum_{y_{(i)} \leq t} \frac{d_i}{n_i}$$

$$\widehat{var}(\tilde{H}(t)) = \sum_{y_{(i)} \leq t} \frac{d_i}{n_i^2}$$

El *estimador de Nelson-Aalen* es la suma acumulada de las probabilidades condicionales estimadas de falla desde el inicio del estudio hasta el final. Este estimador es la primera aproximación de Taylor a (2.10).

2.2.2. El tiempo de supervivencia medio truncado

Un estimador para el tiempo medio de supervivencia truncado sería:

$$\widehat{media} = \int_0^{y_{(n)}} \widehat{S}(t)dt$$

donde $y_{(n)} = max(y_i)$. Pueden ocurrir dos posibilidades para $y_{(n)}$, que sea una observación completa o que sea una censura, en el primer caso, esta integral truncada estaría sobre $[0, \infty]$ puesto que sobre $\left[y_{(n)}, \infty\right]$ el valor de $\widehat{S}(t) = 0$. Pero en cuando $y_{(n)}$ es una censura, el $\lim_{t\to\infty} \widehat{S(t)} \neq 0$, es decir, la integral sobre $[0, \infty]$ estaría indefinida. Para evitar esto, se debe truncar la integral en la última observación, $y_{(n)}$, y se redefine el estimador de K-M como cero después de esta observación. Este estimador es el área total bajo la curva de K-M, y como $\widehat{S(t)}$ es una función discreta, la media se calcula como:

$$\widehat{media} = \sum_{i=1}^{n'} \left(y_{(i)} - y_{(i-1)}\right) \widehat{S}(y_{(i-1)})$$

donde n' = número de observaciones distintas $y_{(i)}$'s, $n' \leq n$, $y_0 = 0$, $\widehat{S}(y_{(i-1)})$ es la altura de la función en $y_{(i-1)}$.

2.3. El modelo de Riesgos Proporcionales de Cox (PH)

El modelo de Cox se escribe en términos de la función de riesgo, que puede tomar cualquier forma, incluso una función escalonada, pero se debe suponer que las funciones de riesgo para diferentes individuos son

proporcionales e independientes del tiempo. También su función de verosimilitud, se vuelve una verosimilitud parcial, y el hecho importante es que la inferencia estadística basada en la función de verosimilitud parcial es similar a aquella basada en la función de verosimilitud, evitando pérdida de información.

2.3.1. Definición del modelo de riesgos proporcionales de Cox

Como su nombre sugiere, el modelo de riesgos proporcionales de Cox posee la propiedad de que los individuos tienen funciones de riesgo que son proporcionales, es decir, $\frac{h(t|\boldsymbol{x}_1)}{h(t|\boldsymbol{x}_2)}$, el cociente de las funciones de riesgo para dos individuos con covariables $\mathbf{x}_1 = (x_{11}, \ldots, x_{p1})'$ y $\mathbf{x}_2 = (x_{12}, x_{22}, \ldots, x_{p2})'$ es una constante (i.e. que no varía con el tiempo t). Esta propiedad implica que la función de riesgo dado un conjunto de covariables $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ se puede escribir como una función de una función de riesgo subyacente y una función $g(x_1, \ldots, x_p)$ de covariables, o sea,

$$h(t|x_1, \ldots, x_p) = h_0(t) \cdot g(x_1, \ldots, x_p)$$

$$h(t|\mathbf{x}) = h_0(t) \cdot g(\mathbf{x}) \tag{2.11}$$

La función de riesgo subyacente, $h_0(t)$, representa el riesgo con el tiempo de los sujetos en la población base o de referencia, y $g(\mathbf{x})$ representa el efecto de las covariables. $h_0(t)$ se puede interpretar como la función de riesgo cuando todas las covariables son cero, es decir, son el vector cero, en cuyo caso $g(\mathbf{x}) = 1$.

El cociente de riesgos de dos individuos con diferentes vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ es

$$\frac{h(t|\mathbf{x}_1)}{h(t|\mathbf{x}_2)} = \frac{h_0(t) \cdot g(\mathbf{x}_1)}{h_0(t) \cdot g(\mathbf{x}_2)} = \frac{g(\mathbf{x}_1)}{g(\mathbf{x}_2)},$$

bajo el supuesto de riesgos proporcionales de este modelo, es una constante, independiente del tiempo.
El modelo de Cox se escribe en términos de la función de riesgo y además supone que la función $g(\mathbf{x})$ es una función exponencial de una combinación lineal de covariables:

$$h(t, \mathbf{X}) = h_0(t) \cdot e^{\sum_{i=1}^{p} \beta_i X_i} \tag{2.12}$$

Definición. Dada una función de riesgo base $h_0(t)$, y valores $x_1, \ldots, x_p$ asociados a un individuo particular, el Modelo de Riesgos Proporcionales de Cox para esa persona está dado por la función de riesgo:

$$h(t|\mathbf{x}) = h_0(t) \times g(\beta_1 x_1 + \cdots + \beta_p x_p) = h_0(t) \times g(\boldsymbol{\beta}'\mathbf{x})$$

donde $g(y) = e^y$ es la función exponencial pues toma sólo valores positivos; $\mathbf{x} = (x_1, \ldots, x_p)'$ es un vector columna de los valores x (llamadas covariables); y $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ es un vector columna de coeficientes.

Se debe señalar que la función de riesgo base es una función de t, pero no de las covariables, en contraste con la expresión exponencial, que involucra las covariables, pero no a t, por lo que se dice que las covariables son independientes del tiempo. Sin embargo, se pueden considerar covariables que dependen de t, las cuales se llaman variables dependientes del tiempo. En este caso el modelo de Cox puede funcionar, y se conoce como el modelo de Cox Extendido que se estudiará más adelante.

2.3.2. La verosimilitud parcial del modelo de Cox

En el modelo de riesgos proporcionales de Cox, los parámetros son las β's en general. Para estimarlos se plantea una función de verosimilitud parcial que se basa en una probabilidad condicional de falla. Los correspondientes estimadores máximo verosímiles para este modelo se denotan por $\widehat{\boldsymbol{\beta}}_{MV}$. Donde estos últimos se obtendrán vía su correspondiente función de verosimilitud, que es una expresión que describe la probabilidad conjunta de obtener los datos observados de los sujetos en

estudio como una función de parámetros desconocidos (en este caso el vector $\boldsymbol{\beta}$) para el modelo considerado.

Procedimiento de estimación del tiempo de supervivencia sin observaciones repetidas

Supóngase que se tienen n observaciones de las cuales k son tiempos de falla, y que las $(n-k)$ restantes presentan censura por la derecha. Sean $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ los tiempos de falla ordenados con covariables correspondientes $\mathbf{x_{(1)}}, \mathbf{x_{(2)}}, \ldots \mathbf{x_{(k)}}$. Sea $\mathbf{R}(t_{(i)})$ el conjunto de riesgo al tiempo $t_{(i)}$, el cual consiste en todos los individuos vivos y no censurados al tiempo $t_{(i)}$. Para la falla en el tiempo $t_{(i)}$, condicionado al conjunto de riesgo $\mathbf{R}(t_{(i)})$, la probabilidad de la falla de este individuo hasta el momento en el que se observó es

$$\frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{j(i)}\right)}{\sum_{l \epsilon \mathbf{R}(t_{(i)})} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)}$$

es decir, toma en cuenta en el denominador a toda la población, incluyendo a los individuos que *fallaron* y a los que resultaron censurados, y el numerador toma en cuenta únicamente a los individuos que fallaron o les ocurrió la falla en $t_{(i)}$. Cada *falla* contribuye con un factor y por eso la verosimilitud parcial es

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left(\sum_{j=1}^{p} \beta_j \cdot x_{j(i)}\right)}{\sum_{l \epsilon \mathbf{R}(t_{(i)})} \exp\left(\sum_{j=1}^{p} \beta_j x_{jl}\right)} \tag{2.13}$$

y la log-verosimilitud parcial es el logaritmo de la verosimilitud parcial

$$\ell(\boldsymbol{\beta}) = \log L(\beta) = \sum_{i=1}^{k}\sum_{j=1}^{p}\beta_j x_{j(i)} - \sum_{i=1}^{k}\log\left[\sum_{l\epsilon\mathbf{R}(t_{(i)})}\exp\left(\sum_{j=1}^{p}\beta_j x_{jl}\right)\right] \tag{2.14}$$

$$= \sum_{i=1}^{k}\left[\boldsymbol{\beta'}\boldsymbol{x}_{(i)} - \log\left(\sum_{l\epsilon\mathbf{R}(t_{(i)})}\exp(\boldsymbol{\beta'}\boldsymbol{x}_l)\right)\right] \tag{2.15}$$

Los estimadores parciales máximo verosimiles (EPMV) $\hat{\beta}$ de β se pueden obtener resolviendo las siguientes ecuaciones simultáneas de derivadas parciales

$$\frac{\partial(\ell(\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}} = 0$$

ó

$$\frac{\partial(\ell(\boldsymbol{\beta}))}{\partial\beta_u} = \sum_{i=1}^{k}[x_{u(i)} - A_{ui}(\boldsymbol{\beta})] = 0 \qquad u = 1, 2, \ldots, p$$

donde

$$A_{ui}(\beta) = \frac{\sum_{l\epsilon\mathbf{R}(t_{(i)})} x_{ul}\exp(\sum_{j=1}^{p}\beta_j x_{jl})}{\sum_{l\epsilon\mathbf{R}(t_{(i)})}\exp(\sum_{j=1}^{p}\beta_j x_{jl})} = \frac{\sum_{l\epsilon\mathbf{R}(t_{(i)})} x_{ul}\exp(\boldsymbol{\beta'}\mathbf{x_l})}{\sum_{l\epsilon\mathbf{R}(t_{(i)})}\exp(\boldsymbol{\beta'}\mathbf{x}_l)}$$

y por último se aplica el método iterativo de Newton-Raphson. Después se obtienen las segundas derivadas parciales de $\ell(\beta)$ con respecto a β_u y β_v, con $u, v = 1, 2, \ldots, p$, y éstas son

$$I_{uv}(\boldsymbol{\beta}) = \frac{\partial^2\ell(\beta)}{\partial\beta_u\partial\beta_v} = -\sum_{i=1}^{k}C_{uvi}(\beta_1, \ldots, \beta_p) \tag{2.16}$$

$$= -\sum_{i=1}^{k} C_{uvi}(\beta) \qquad u, v = 1, 2, \ldots, p$$

donde

$$C_{uvi}(\beta) = \frac{\sum_{l \epsilon \mathbf{R}(t_{(i)})} x_{ul} x_{vl} \exp(\sum_{j=1}^{p} \beta_j x_{jl})}{\sum_{l \epsilon \mathbf{R}(t_{(i)})} \exp(\sum_{j=1}^{p} \beta_j x_{jl})} - A_{ui}(\beta) A_{vi}(\beta) \quad (2.17)$$

La matriz de covarianzas estimadas del EPMV $\widehat{\boldsymbol{\beta}}$, se denota como $\widehat{Var}(\hat{\boldsymbol{\beta}})$ o $\widehat{Cov}(\hat{\boldsymbol{\beta}})$ y es:

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \widehat{Cov}(\widehat{\boldsymbol{\beta}}) = \left[-\frac{\partial^2 l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right]^{-1}$$

donde $-\frac{\partial^2 l(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$ se conoce como la *matriz de información observada* con $-I_{uv}(\widehat{\boldsymbol{\beta}})$ como su elemento (u, v) y donde $I_{uv}(\boldsymbol{\beta})$ se define anteriormente por la ecuación (2.16). Sea el elemento $v_{(i,j)}$ de $\widehat{Var}(\widehat{\boldsymbol{\beta}})$ en la ecuación, entonces el $100 \times (1 - \alpha)\,\%$ intervalo de confianza para β_i es:

$$\left(\widehat{\beta}_i - z_{\alpha/2} \sqrt{v_{ii}}, \quad \widehat{\beta}_i + z_{\alpha/2} \sqrt{v_{ii}} \right)$$

donde $z_{\alpha/2}$ es el cuantil $\alpha/2$ de una variable aleatoria normal estándar. Puesto que el vector de parámetros estimados se distribuye asintóticamente normal, y con ello también cada parámetro estimado.

Procedimiento de estimación con observaciones del tiempo de supervivencia repetidas

De entre n tiempos de supervivencia observados se tienen k diferentes tiempos de falla $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$. Sea $m_{(i)}$ el número de personas

que fallan en $t_{(i)}$; $m_{(i)} > 1$ si hay más de una observación con valor $t_{(i)}$; $m_{(i)} = 1$ si sólo hay una observación con valor $t_{(i)}$. Sea $R(t_{(i)})$ el conjunto de individuos en riesgo al tiempo $t_{(i)}$, i.e. $R(t_{(i)})$ consiste de aquellos tiempos de supervivencia que son al menos $t_{(i)}$, y r_i es el número de esos individuos.

Para tener certeza con las observaciones repetidas, se agrega la siguiente notación. De cada conjunto $R(t_{(i)})$, se pueden seleccionar $m_{(i)}$ sujetos aleatoriamente que se denotan por $\mathbf{u}_{(j)}$. Hay

$$\binom{r_i}{m_{(i)}} = \frac{r_i!}{[m_{(i)}!(r_i - m_{(i)})!]}$$

posibles $\mathbf{u}_{(i)}$'s. Sea $\mathbf{U}_i$ el conjunto que contiene a todas las $\mathbf{u}_{(i)}$'s. Ahora estudiando las observaciones empatadas, sean $\mathbf{x_k} = (x_{1k}, x_{2k}, \ldots, x_{pk})'$ las covariables del k-ésimo indiviuo,

$$z_{\mathbf{u}(j)} = \sum_{k \epsilon \mathbf{u}_{(j)}} \mathbf{x}_k = (z_{1\mathbf{u}(j)}, z_{2\mathbf{u}(j)}, \ldots, z_{p\mathbf{u}(j)})'$$

donde $z_{l\mathbf{u}(j)}$ es la suma de la l-ésima covariable de las $m_{(i)}$ personas quienes están en $\mathbf{u}_{(j)}$. Sea $\mathbf{u}^*_{(i)}$ el conjunto de $m_{(i)}$ personas que fallaron al tiempo $t_{(i)}$, y

$$\mathbf{z}_{u^*(i)} = \sum_{k \epsilon \mathbf{u}^*_{(i)}} \mathbf{x}_k = (z_{1\mathbf{u}^*(i)}, z_{2\mathbf{u}^*(i)}, \ldots, z_{p\mathbf{u}^*(i)})'$$

donde $z^*_{l\mathbf{u}^*(i)}$ es la suma de la l-ésima covariable de las $m_{(i)}$ personas que están en $\mathbf{u}^*_{(i)}$ (que fallaron al tiempo $t_{(i)}$).

Escala de tiempo continuo. En el caso de una escala de tiempo continuo, para las $m_{(i)}$ personas que fallan en $t_{(i)}$, es razonable decir que los tiempos de supervivencia de las $m_{(i)}$ personas no son idénticos puesto que los empates son en su mayoría resultado de mediciones imprecisas. Si se

pudieran hacer mediciones precisas, estos $m_{(i)}$ tiempos de supervivencia podrían ser ordenados y se podría usar la función de verosimilitud como en la ecuación (2.13). En la ausencia de conocimiento del verdadero orden (el caso real), se tienen que considerar todas las posibles ordenaciones de estos $m_{(i)}$ tiempos repetidos de supervivencia. Para cada $t_{(i)}$, los $m_{(i)}$ tiempos repetidos de supervivencia se pueden ordenar en $m_{(i)}!$ maneras diferentes. Para cada uno de estos posibles ordenamientos se tiene un producto como en la ecuación (2.13) para los correspondientes $m_{(i)}$ tiempos. Por eso, cuando el tiempo de supervivencia se mide en una escala de tiempo continuo, la construcción y el cálculo de la función de verosimilitud parcial se vuelve una tarea tediosa para $m_{(i)}$'s grandes.

Para aproximar la función de verosimilitud parcial exacta, las siguientes dos funciones se pueden usar cuando $m_{(i)}$ es grande comparada con r_i. La primera aproximación la publicó Breslow en 1974:

$$L_B(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left(\mathbf{z}'_{\boldsymbol{u}^*(i)}\boldsymbol{\beta}\right)}{\left[\sum_{l\epsilon R(t_{(i)})} \exp\left(\mathbf{x}'_l\boldsymbol{\beta}\right)\right]^{m_{(i)}}} \tag{2.18}$$

la segunda aproximación fue desarrollada por Efron en 1977:

$$L_E(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left(\mathbf{z}'_{\boldsymbol{u}^*(i)}\boldsymbol{\beta}\right)}{\prod_{j=1}^{m_{(i)}} \left[\sum_{l\epsilon R(t_{(i)})} \exp\left(\boldsymbol{x}'_l\boldsymbol{\beta}\right) - \left[\frac{(j-1)}{m_{(i)}}\right] \sum_{l\epsilon \mathbf{u}^*_{(i)}} \exp(\mathbf{x}'_l\boldsymbol{\beta})\right]} \tag{2.19}$$

Escala de tiempo discreto. Si los tiempos de supervivencia son observados en tiempo discreto, las observaciones reetidas son empates reales, es decir, estos eventos sí ocurren al mismo tiempo. Cox en 1972 publicó el

siguiente modelo logístico:

$$\frac{h_i(t)dt}{1-h_i(t)dt} = \frac{h_0(t)dt}{1-h_0(t)dt}\exp\left(\sum_{j=1}^{p}\beta_j x_{jl}\right) = \frac{h_0(t)dt}{1-h_0(t)dt}\exp\left(\boldsymbol{\beta}'\mathbf{x}_i\right)$$

Este modelo se reduce a (2.11) en la escala de tiempo continuo. Usando el modelo y reemplazando el i-ésimo término en (2.13) con el siguiente término con observaciones repetidas en $t_{(i)}$:

$$\frac{\exp \mathbf{z}_{\mathbf{u}^*}(i)'\boldsymbol{\beta}}{\sum_{\mathbf{u}_{(j)}\in\mathbf{U}_i}\exp \mathbf{z}'_{\mathbf{u}(j)}\boldsymbol{\beta}}$$

la función de verosimilitud parcial con observaciones empatadas en escala de tiempo discreto es:

$$L_d(\mathbf{d}) = \prod_{i=1}^{k}\frac{\exp \mathbf{z}'_{\mathbf{u}^*(i)}\boldsymbol{\beta}}{\sum_{\mathbf{u}_{(j)}\in\mathbf{U}_i}\exp \mathbf{z}'_{\mathbf{u}(j)}\boldsymbol{\beta}} \tag{2.20}$$

El i-ésimo término en esta expresión representa la probabilidad condicional de observar las $m_{(i)}$ fallas dado que hay $m_{(i)}$ fallas en el tiempo $t_{(i)}$ y el conjunto de riesgo $R(t_{(i)})$ en $t_{(i)}$. El número de términos en el denominador del i-ésimo término es $\binom{r_i}{m_{(i)}}$, como se dijo antes, y será muy grande si los $m_{(i)}$ son grandes. Afortunadamente, hay un algoritmo recursivo desarrollado por Gail (1981) que hace los cálculos mucho más manejables. La ecuación (2.20) se considera como una aproximación de la función de verosimilitud parcial para tiempos de supervivencia continuos con empates, suponiendo que los empates son válidos, como si fuesen observados en la escala de tiempo discreto.

En la mayoría de las situaciones prácticas, las tres funciones de verosimilitud parcial que se acaban de mencionar son una aproximación razonablemente buena de la función de verosimilitud parcial exacta para tiempos de supervivencia continuos y con empates. Cabe mencionar que

cuando no hay empates en los tiempos de evento (i.e., $m_{(i)} \equiv 1$), (2.18), (2.19) y (2.20) se reducen a (2.13). El estimador parcial máximo verosímil de $\boldsymbol{\beta}$ en (2.18), (2.19) y (2.20) se puede estimar con procedimientos similares a los que se presentan cuando no hay empates en la subsección anterior.

2.4. Estimación de la función de supervivencia

La funcíon de supervivencia se puede estimar con el uso de modelos paramétricos de regresión, simplemente reemplazando los parámetros y coeficientes en la función de supervivencia con sus estimadores. Sin embargo, esto no funciona para el modelo de riesgos proporcionales de Cox, pues no se conoce exactamente la forma de la función de riesgo base, y por ende, tampoco la de supervivencia. Por lo que se usa una alternativa de dos estimadores de la función de supervivencia que propusieron Breslow (1974) y Kalbfleisch y Prentice (1980). Estos estimadores se pueden calcular por medio de software estadístico.

A continuación se calcula la función de supervivencia bajo el modelo de Cox, con covariables $x_j's$ como:

$$S(t,\mathbf{x}) = \exp\left(-\int_0^t h(u|\mathbf{x})du\right) = \exp\left(-\exp(\mathbf{x}'\boldsymbol{\beta})\int_0^t h_0(u)du\right) \tag{2.21}$$

$$= \left(\exp\left(-\int_0^t h_0(u)du\right)\right)^{\exp(\boldsymbol{x}'\boldsymbol{\beta})} = [S_0(t)]^{\exp(\mathbf{x}'\boldsymbol{\beta})}$$

donde $S_0(t)$ es la función de supervivencia base. Asimismo se puede calcular la función de densidad de probabilidad de T como:

$$f(t,\mathbf{x}) = h_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})[S_0(t)]^{\exp(\mathbf{x}'\boldsymbol{\beta})} \tag{2.22}$$

Hay dos generalizaciones importantes:

- La función de riesgo base $h_0(t)$ puede variar en subconjuntos específicos de los datos.
- Las variables de regresión $\mathbf{x}$ pueden depender del tiempo, es decir, $\mathbf{x} = \mathbf{x}(t)$.

Una vez que los coeficientes de regresión, los $\beta's$, son estimados, sólo faltaría obtener la función de supervivencia base $S_0(t)$. De la función de supervivencia estimada se puede calcular la probabilidad de sobrevivir más que un determinado tiempo para algún individuo con covariables $x_1, \ldots, x_p$.

Bajo el supuesto de que la función de riesgo base es constante entre cada par observado de tiempos sucesivos de falla, Breslow (1974) propuso el siguiente estimador de la función de riesgo acumulado base

$$\widehat{H}_0(t) = \sum_{t_{(i)} \leq t} \frac{m_{(i)}}{\sum_{l \in R_{t_{(i)}}} \exp(\mathbf{x}_l' \widehat{\boldsymbol{\beta}})}$$

Siguiendo la ecuación, la función de supervivencia base se puede estimar como

$$\widehat{S}_0(t) = \exp[-\widehat{H}_0(t)] = \prod_{t_{(i)} \leq t} \left\{ \exp \left[\frac{m_{(i)}}{\sum_{l \in R_{t_{(i)}}} \exp(\mathbf{x}_l' \widehat{\boldsymbol{\beta}})} \right] \right\}$$

y la función de supervivencia del modelo de Cox para un individuo con un conjunto de covariables $\mathbf{x} = (x_1, \ldots, x_p)$ está dado por

$$\widehat{S}(t; \mathbf{x}) = [\widehat{S}_0(t)]^{\exp(\sum_{j=1}^{p} \widehat{\beta}_j x_j)} = [\hat{S}_0(t)]^{\exp(\widehat{\boldsymbol{\beta}}' \mathbf{x})}$$

$\widehat{S}(t; \mathbf{x})$ tiene una distribución asintóticamente normal con media $S(t, \mathbf{x})$. Como $S(t; \mathbf{x}) = \exp[-H(t, \mathbf{x})]$, un estimador de la varianza $\widehat{Var}(\widehat{S}(t, \mathbf{x}))$ de $\widehat{S}(t, \mathbf{x})$, calculado con el método delta es:

$$\widehat{Var}(\widehat{S}(t,\mathbf{x})) \simeq [\widehat{S}(t,\mathbf{x})]^2 \widehat{Var}(\widehat{H}(t,\mathbf{x}))$$

No se obtendrá $\widehat{H}(t,\mathbf{x})$ dada su complejidad. Las bandas de confianza asintóticas para la función de supervivencia son:

$$\left(\widehat{S}(t,\mathbf{x}) - Z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{S}(t,\mathbf{x}))}, \widehat{S}(t,\mathbf{x}) + Z_{\alpha/2}\sqrt{\widehat{Var}(\widehat{S}(t,\mathbf{x}))}\right)$$

donde $z_{\alpha/2}$ es el $100 \times (1-\alpha/2)$ cuantil superior de una distribución normal estándar.

2.5. Diagnóstico y evaluación del modelo de riesgos proporcionales

Cualquier modelo estadístico de datos de supervivencia se debe validar y evaluar, pues para identificar riesgos y factores de pronóstico importantes es necesario revisar que el modelo sea adecuado, de otro modo los resultados arrojados por éste serán imprecisos. Esta es una parte esencial de la estadística, que por supuesto no excluye al modelo de riesgos proporcionales.

2.5.1. Residuos para Bondad de Ajuste

Ya se han desarrollado varios métodos para enfrentar la situación de evaluar los residuos, para diversos ajustes, el más popular y conocido es el que se hace en regresión lineal, donde se tiene una medida de discrepancia de los valores observados contra los valores ajustados, que en comparación con las técnicas de supervivencia, es más sencillo. En el análisis de supervivencia se tiene que lidiar con la censura, lo cual complica la evaluación de los residuos.

A continuación se presentan tres tipos de residuos para el modelo de

riesgos proporcionales, los cuales pueden graficarse contra el tiempo de supervivencia o contra una covariable y con ello obtener información sobre qué tan apropiado es el modelo de riesgos proporcionales. También proporcionan información sobre los comportamientos atípicos de algunas observaciones. Asimismo, como en muchas de las evaluaciones y calificaciones de modelos y ajustes, siempre puede haber subjetividad en la interpretación de los residuos.

Residuo de Cox-Snell (Cox y Snell, 1968)

Este residuo se obtendrá a partir del siguiente teorema, acerca de la distribución de una variable aleatoria:

Teorema 2.5.1. *Sea T la variable aleatoria del tiempo de supervivencia y sea $Y = -\log(S(T))$ donde $S(T)$ es su función de supervivencia, entonces*

$$Y \sim \exp(1)$$

Demostración. Utilizando el teorema de cambio de variable en el caso univariado, se tiene que si $f_X(x)$ es la función de densidad de una variable aleatoria X, entonces la función de densidad de $Y = g(X)$ está dada por

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left(\left| \frac{dy}{dx} \right| \right)^{-1} \tag{2.23}$$

donde $f_X(x)$ es la densidad de X expresada en términos de y. Ahora, sea $Y = -\log(S(T)) \quad \Longrightarrow S(t) = e^{-y} \quad \Longrightarrow t = S^{-1}(e^{-y})$ Sustituyendo en (2.23) la inversa, se tiene que

$$f_Y(y) = \frac{f_T(S^{-1}(e^{-y}))}{|\frac{dy}{dt}|}$$

pero

$$\frac{dy}{dt} = \frac{d}{dt}\left[-\log(S(t))\right] = \frac{f_T(t)}{S_T(t)}$$

y evaluando la inversa, se tiene

$$\frac{f_T(S_T^{-1}(e^{-y}))}{S_T(S_T^{-1}(e^{-y}))} = \frac{f_T(S^{-1}(e^{-y}))}{e^{-y}}$$

Y finalmente, sustituyendo

$$f_Y(y) = \frac{f_T(S^{-1}(e^{-y}))}{\left|\frac{f_T(S^{-1}(e^{-y}))}{e^{-y}}\right|} = e^{-y}$$

Por lo tanto

$$Y \sim exp(1)$$

Una vez demostrado el teorema, es claro, que el modelo ajustado a los datos observados será satisfactorio si los valores del estimador de $\widehat{S}(t_i)$ para el i-ésimo individuo en t_i son muy cercanos a los de $S(t_i)$, es decir, el estimador se comportará como n observaciones de una distribución exponencial censurada con media 1.

Los residuos de *Cox-Snell* para el modelo de riesgos proporcionales, r_{C_i} para el individuo i con tiempo observado de supervivencia t y covariables $\mathbf{x}_i$ se definen como

$$r_{c_i} = \widehat{H}_0(t_i) \cdot e^{\widehat{\boldsymbol{\beta}} X_i} = \widehat{H}_i(t_i) = -\log \widehat{S}(t_i; \mathbf{x}_i)$$

los cuales son los riesgos acumulados estimados basados en el modelo de riesgos proporcionales. Si el tiempo t_i observado es censurado, también lo será el respectivo r_{c_i}. Para saber si el modelo de riesgos proporcionales es adecuado, la gráfica de r_{c_i} y su función de supervivencia estimada por Kaplan-Meier ($\hat{S}_{KM}(r_{c_i})$) tendría una línea recta a un ángulo de 45 grados.

Residuo de martingalas (Therneau et al. 1990)

Haciendo una modificación a los residuos de *Cox-Snell*, se puede obtener nuevos residuos que tendrían media unitaria para observaciones no

censuradas. Éstos pueden ser más refinados haciendo un reescalamiento para que tengan media cero cuando hay una observación censurada. Si además los valores resultantes se multiplican por -1, se obtienen los *residuos de martingalas*

$$r_{Mi} = \delta_i - r_{c_i} \qquad i = 1, 2, \ldots, n \tag{2.24}$$

Estos residuos, utilizan r_{c_i} basados en el estimador de *Nelson-Aalen* de la función de riesgo acumulado.

En muestras grandes son no correlacionados y tienen valor esperado igual a cero, pero no son simétricamente distribuidos alrededor de cero. Los *residuos de martingalas* son siempre negativos para observaciones censuradas y se utilizan frecuentemente para evaluar la bondad de ajuste del modelo de riesgos proporcionales.

Los *residuos de martingalas* se interpretan como la diferencia entre el número observado de fallas para un individuo en el intervalo $(0, t)$ denotado por δ_i y el número esperado de fallas de acuerdo al modelo. r_{M_i} se grafica contra el orden de rango de tiempo. Idealmente, no debería exhibir ningún comportamiento si el modelo es adecuado.

Residuos de devianza

Aunque los *residuos martingala* comparten muchas de las propiedades de los residuos que se manejan en otras situaciones, tales como los del análisis regresión lineal, no son simétricamente distribuidos sobre cero, incluso cuando el modelo ajustado es el correcto. Los residuos de devianza, que fueron desarrollados por Therneau *et.al* (1990), se distribuyen simétricamente sobre cero. Y se definen como

$$r_{Di} = sign(R_{Mi})\sqrt{2[-R_{Mi} - \delta_i \log(\delta_i - R_{Mi})]} \qquad i = 1, 2, \ldots, n \tag{2.25}$$

donde $sign()$ es la función signo, la cual toma valores 1 si el argumento es positivo, 0 si es cero, y -1 si es negativo.

Los *residuos de devianza* son *residuos de martingalas* que han sido transformados para producir valores que se distribuyen simétricamente alrededor cero cuando el modelo ajustado es el apropiado.

Es sencillo ver que los *residuos de martingala* r_{Mi} pueden tomar cualquier valor en el intervalo $(-\infty, 1)$, y para valores grandes negativos de r_{Mi}, el término en los corchetes en la ecuación 2.25 es dominado por r_{Mi}. Tomando la raíz cuadrada de esta cantidad se obtiene el efecto de acercarse mucho más a cero. Si se consideran los residuos de martingalas en el intervalo $(0, 1)$, el término $\delta_i \log(\delta_i - r_{Mi})$ en la ecuación (2.25) será no negativo únicamente para observaciones no censuradas, y tendrá el valor $\log(1 - r_{Mi})$. Mientras más se aproxima r_{Mi} a uno, $\log(q - r_{Mi})$ toma valores negativos grandes. Y con ello los residuos de devianza son expandidos hacía $+\infty$ mientras el residuo de martingalas alcanza su límite superior de uno.

Residuos de Schoenfeld

Los residuos que se acaban de describir, tienen dos desventajas. La dependencia que tienen en los tiempos obervados de supervivencia y además de que requieren un estimador de la función de riesgo acumulado. Ambas desventajas se pueden despreciar, con la propuesta que realizó Schoenfeld (1982), puesto que los *residuos de Schoenfeld* no usan un valor para cada individuo, pero sí, un conjunto de valores, uno para cada variable explicativa incluida en el ajuste del modelo de riesgos proporcionales de Cox.

De este modo, el i-ésimo residuo de Schoenfeld para X_j, la j-ésima variable explicativa en el modelo, está dado por

$$r_{Sji} = \delta_i\{x_{ji} - \hat{a}_{ji}\}$$

donde x_{ji} es el valor de la j-ésima variable explicativa $j = 1, 2, \ldots, p$ para el i-ésimo individuo en el estudio,

$$\hat{a}_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_l)}$$

y $R(t_i)$ es el conjunto de todos los individuos en riesgo al tiempo t_i.

Nótese que los valores distintos de cero de estos residuos son solamente para observaciones no censuradas. Más aún, si la observación más grande en una muestra de tiempos de supervivencia, es no censurada, el valor de $\hat{a}_{ji}$ para esa obervación, de la ecuación 2.5.1, será igual a x_{ji} y por ello $r_{Pij} = 0$. Para distinguir los residuos que son genuinamente cero de aquellos que se obtienen de observaciones censuradas, los últimos se expresan como valores faltantes.

El i-ésimo *residuo de Schoenfeld*, para la variable explicativa X_j, es un estimador del i-ésimo componente de la primera derivada del logaritmo de la función de verosimilitud parcial con respecto a β_j, la cual, de la ecuación (2.14), está dado por

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i\{x_{ji} - a_{ji}\}$$

donde

$$a_{ji} = \frac{\sum_{\ell} x_{jl} \exp(\boldsymbol{\beta}' \mathbf{x}_l)}{\sum_{\ell} \exp(\boldsymbol{\beta}' \mathbf{x}_\ell)}$$

el i-ésimo término en la suma, evaluado en $\hat{\boldsymbol{\beta}}$, es entonces el *residuo de Schoenfeld* para X_j dado en la ecuación (2.5.1). Y como los estimadores

de los β's son tales que

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j}|_{\hat{\boldsymbol{\beta}}} = 0$$

los *residuos de Schoenfeld* deben sumar cero. Estos residuos también tienen la propiedad de que, para muestras grandes, el valor esperado de r_{Pji} es cero, y son no correlacionados unos con otros.

2.6. El Modelo de Riesgos Proporcionales de Cox para variables dependientes del tiempo

En la práctica es común que un estudio incluya ambos tipos de covariables, tanto dependientes como independientes del tiempo. Por ejemplo, algunas variables como el género y edad (si hablamos de un periodo menor a doce meses), no cambian con el tiempo y son recolectadas generalmente solamente una vez en un estudio. Existen otras variables que pueden variar con el tiempo y frecuentemente se recolectan en diversos puntos del tiempo. Es por esto que se ve la necesidad de modificar un poco la función de verosimilitud (2.13), para poder admitir covariables que varíen con el tiempo. Algunos ejemplos de estas covariables son: el tiempo en el empleo; donde se indica si un individuo continúa con empleo. Estatus de fumador o niveles de obesidad al tiempo t.
De este modo, la verosimilitud parcial quedaría entonces:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\left[\sum_{j=1}^{p} \beta_j x_{j(i)}(t_{(i)})\right]}{\sum_{l \epsilon \mathbf{R}(t_{(i)})} \exp\left[\sum_{j=1}^{p} \beta_j x_{jl}(t_{(i)})\right]}$$

$$= \sum_{i=1}^{k} \frac{\exp\left[\boldsymbol{\beta}' \mathbf{x}_{j(i)}(t_{(i)})\right]}{\sum_{l \epsilon \mathbf{R}(t_{(i)})} \exp\left[\boldsymbol{\beta}' \mathbf{x}_{jl}(t_{(i)})\right]} \tag{2.26}$$

donde k es el número de tiempos de falla distintos, $R(t_{(i)})$ es el conjunto de riesgo que contiene todas las personas en riesgo al tiempo $t_{(i)}$, $\mathbf{x}_l(t_{(i)}) = (x_{1l}(t_{(i)}), x_{2l}(t_{(i)})), \ldots, x_{pl}(t_{(i)}))'$ denota las covariables observadas de la persona ℓ en el evento ordenado no censurado del tiempo $t_{(i)}$, y $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)'$ denota los coeficientes desconocidos. Por supuesto que para las variables que no varían con el tiempo sus valores permanecen constantes en el tiempo. Por ejemplo, sea x_{1k} la variable asignada para el género de la persona k, entonces $x_{1k}(t) = x_{1k}(0) = x_{1k}$ para toda t. Por lo tanto, en la práctica la verosimilitud tiene combinaciones de covariables que dependen y que no dependen del tiempo. La manera de estimar los coeficientes es muy parecida al método que se usó en la sección anterior, además de que los paquetes computacionales como R, facilitan bastante el trabajo.

Existen dos tipos de covariables dependientes del tiempo: las primeras son observadas repetidamente en diferentes puntos del tiempo previo a la ocurrencia del evento o al final del estudio o el tiempo de censura, y las segundas son aquellas covariables que cambian con el tiempo de acuerdo a una función matemática y covariables que tienen diferentes valores debido a diferentes factores o condiciones.

Capítulo 3

Aplicación: datos de deudas

Los sistemas del *credit scoring* tradicionalmente buscan estimar la probabilidad de que un individuo deje de pagar un cierto número de veces (previamente establecidas) encontrándolo en una situación de *default*, es decir, de acuerdo a la definición de malo previamente establecida. Sin embargo, utilizando el análisis de *supervivencia*, será más importante en qué momento del tiempo, el cliente tendrá *default*.

Este último capítulo tiene como finalidad describir el funcionamiento de un *credit scoring*, utilizando el modelo de riesgos proporcionales de Cox, y así culminar con la integración de los dos temas tratados en los capítulos anteriores.

Se utilizó una base de $1,225$ clientes de una entidad prestataria, esta entidad, es en cierto modo ficticia puesto que los datos de los clientes han sido modificados. Los datos contienen información sobre clientes que adquirieron un préstamo personal sin garantía con un plazo fijo mensual de 36 meses. La base de datos se refiere a diversas características de clientes potenciales para la adquisición de un préstamo, dichas características, se describen a continuación, enunciando el tipo de variable, ya sea numérica o categórica:

No.	Característica	Tipo	
1	x_1	Edad del cliente	Numérica
2	x_2	Número de hijos	Numérica
3	x_3	Número de otros dependientes	Numérica
4	x_4	Tiene teléfono en casa	Categórica
5	x_5	Salario del conyuge	Numérica
6	x_6	Estatus del solicitante	Categórica
7	x_7	Salario del solicitante	Numérica
8	x_8	Estatus residencial	Categórica
9	x_9	Valor de la casa	Numérica
10	x_{10}	Balance de la hipóteca	Numérica
11	x_{11}	Gastos en hipóteca o renta	Numérica
12	x_{12}	Gastos en deudas	Numérica
13	x_{13}	Gastos en ventas a plazos	Numérica
14	x_{14}	Gastos en tarjetas de crédito	Numérica
15	x_{15}	Término de la deuda	Categórica
16	x_{16}	Indicador *bueno/malo*	Categórica
17	x_{17}	Tiempo al *Default*	Numérica
18	x_{18}	Tiempo al *pago temprano*	Numérica

Cabe mencionar que cuando se utiliza una base de datos de una entidad prestataria, la cantidad de datos e información de variables que se tiene es muy grande, con lo que los costos de procesamiento y verificación requieren de software y equipo de gran capacidad. Esta cantidad de datos ayuda a tener una mejor discriminación y una selección mucho más exhaustiva de variables que además interactúan con otras, llegando a una clasificación ordinaria mucho más predictiva por lo que la base que se utiliza en este trabajo se ve muy limitada, sin embargo, a continuación se describe el bosquejo del desarrollo de una *scorecard*.

3.1. Metodología

Sea T el tiempo hasta caer en *default* o *pago temprano*. Sea T_1 el tiempo hasta el *default*, y de la misma forma, pero separadamente, se define T_2 el tiempo hasta el *pago temprano*. Así, se aplicará el *análisis de supervivencia* de manera separada a cada una de estas variables T_1 y T_2. Entonces, de acuerdo a Thomas (1999), el tiempo de vida de la deuda se puede estimar como el $min\{T_1, T_2\}$.

Se describirá el comportamiento con el modelo de Cox, tomando en cuenta primeramente la característica de *default* como tiempo de falla. Se hará el análisis estadístico pertinente como se describe en la parte de análisis de supervivencia.

El tratamiento de cada variable es distinto según su tipo, ya sea numérica o categórica, funcionarán como covariables en el modelo de Cox. Seguido de esto, se creará la *scorecard* a través del riesgo que representen las variables, puesto que, como es de esperarse, algunas tendrán mayor riesgo que otras.

3.2. Análisis descriptivo de supervivencia

Para conocer mejor los datos con los que se trabajan y como buena costumbre estadística, se realiza un análisis descriptivo. Primeramente se muestra una tabla en donde se describen cuatro tipos de variable en la base de datos, y se específica la tasa de aprobación por cada característica:

	Total	Malos (1)	Buenos (0)	Tasa de aprobación
Default	508	133	375	74 %
Early Repayment	216	59	157	73 %
Paid off	330	93	237	72 %
Still open	171	38	133	78 %
	1225	323	902	74 %

De donde es evidente que se llevó a cabo una buena segmentación para la mayoría de las características, sin embargo, también cabe mencionar que las proporciones para cada característica con respecto a cada variable varían como se muestra en la siguiente tabla:

	Total	Tasa por característica
Default	508	41 %
Early Repayment	216	18 %
Paid off	330	27 %
Still open	171	14 %
	1225	

A continuación se realiza un análisis que permitirá conocer cómo se comportan los tiempos de falla y la censura de la base de datos de los clientes que ya recibieron un préstamo. Para ello se ha utilizado R y la paquetería OISurv.

En resumen para los tiempos de falla considerados como *default* se tiene lo siguiente:

```
Call: survfit(formula = my.survd ~ 1)

   records       n.max     n.start      events      *rmean
  1225.000    1225.000    1225.000     508.000      27.730
  *se(rmean)   median   0.95LCL     0.95UCL
    0.277      34.000   27.000           NA
```

```
    * restricted mean with upper limit =  36
```

time	n.risk	n.event	survival	std.err	lower 95 % CI	upper 95 % CI
7	1216	1	0.999	0.000822	0.998	1
8	1195	1	0.998	0.001172	0.996	1
10	1139	1	0.997	0.001462	0.995	1
11	1109	5	0.993	0.002479	0.988	0.998
12	1074	16	0.978	0.004409	0.97	0.987
13	1034	10	0.969	0.005284	0.958	0.979
14	1003	29	0.941	0.007253	0.927	0.955
15	964	35	0.907	0.008998	0.889	0.924
16	921	35	0.872	0.010371	0.852	0.893
17	879	46	0.826	0.011811	0.804	0.85
18	832	40	0.787	0.012806	0.762	0.812
19	791	53	0.734	0.013844	0.707	0.762
20	737	53	0.681	0.014625	0.653	0.711
21	684	43	0.638	0.015093	0.609	0.669
22	641	36	0.603	0.015383	0.573	0.633
23	605	18	0.585	0.015495	0.555	0.616
24	587	24	0.561	0.015611	0.531	0.592
25	563	15	0.546	0.015664	0.516	0.577
26	548	10	0.536	0.015692	0.506	0.567
27	538	10	0.526	0.015713	0.496	0.558
28	528	7	0.519	0.015724	0.489	0.551
29	521	3	0.516	0.015728	0.486	0.548
30	518	4	0.512	0.015732	0.482	0.544
31	514	7	0.505	0.015737	0.475	0.537
32	507	1	0.504	0.015737	0.474	0.536
33	506	2	0.502	0.015738	0.472	0.534
34	504	3	0.499	0.015738	0.469	0.531

```
Call: survfit(formula = my.survpe ~ 1)
```

```
   records       n.max    n.start      events      *rmean
  1225.000   1225.000   1225.000    216.000      31.452
*se(rmean)      median    0.95LCL    0.95UCL
     0.282          NA         NA          NA
    * restricted mean with upper limit =  36
```

time	n.risk	n.event	survival	std.err	lower 95 % CI	upper 95 % CI
5	1225	1	0.999	0.000816	0.998	1
6	1224	8	0.993	0.00244	0.988	0.997
7	1216	20	0.976	0.004344	0.968	0.985
8	1195	26	0.955	0.005919	0.944	0.967
9	1168	29	0.931	0.007227	0.917	0.946
10	1139	29	0.908	0.008276	0.892	0.924
11	1109	30	0.883	0.009187	0.865	0.901
12	1074	24	0.863	0.009825	0.844	0.883
13	1034	21	0.846	0.010344	0.826	0.866
14	1003	10	0.837	0.010579	0.817	0.858
15	964	8	0.83	0.010772	0.81	0.852
16	921	7	0.824	0.010951	0.803	0.846
17	879	1	0.823	0.010979	0.802	0.845
18	832	1	0.822	0.01101	0.801	0.844
19	791	1	0.821	0.011045	0.8	0.843

Haciendo un análisis por inspección de los 1225 clientes, y utilizando los estimadores descritos en el capítulo anterior, se destaca la ocurrencia de 508 *fallas* y 717 censuras, es decir, 508 personas cayeron en *default* y el resto se censuraron, también se tiene una mediana de 34 meses, la cual se puede contrastar con la media que es de 27 meses y está truncada en $t = 36$ puesto que es el último tiempo que se toma en consideración. Asimismo, la variable *pago temprano*, arrojó 216 *fallas*, con una media de 31 meses también truncada en $t = 36$.

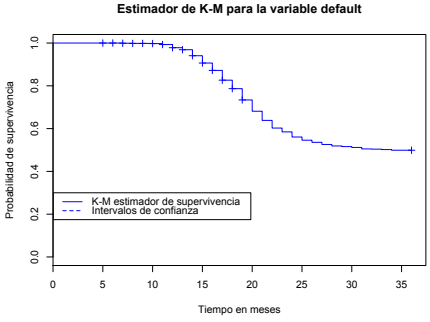

Figura 3.1: *Estimador K-M para la variable default.*

De inmediato se pueden contrastar también las gráficas de la función de supervivencia generadas por ambas variables, en donde el comportamiento, como era de esperarse, tiene mayores probabilidades de sobrevivir al *pago temprano* que sobrevivir al *default* conforme pasa el tiempo, es decir, que es más probable que una persona deje de pagar a que pague antes de tiempo, coincidiendo con la media y mediana de los tiempos de supervivencia antes mencionados (27 y 31 meses).

3.3. Tratamiento de las características

Se estudiaron todas las características y como ejemplo se revisará el ingreso del solicitante con un pequeño resumen para evaluar la variable y verificar impactos. Primero se observa la tabla con un resumen descriptivo de la variable ingreso:

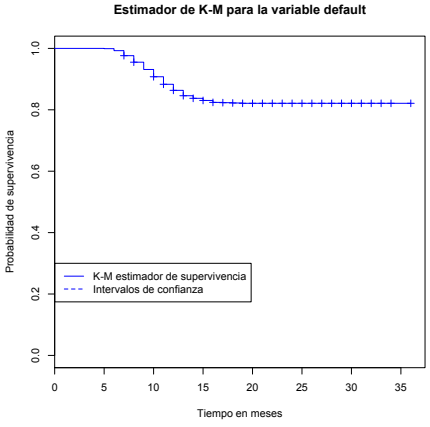

Figura 3.2: *Estimador K-M para la variable pago temprano.*

Resumen descriptivo	
Media	$\$21,240$
Mediana	$\$19,500$
Varianza	$\$252,689,428$
Desviación estándar	$\$15,896.21$
Primer cuartil	$\$9,000$
Tercer cuartil	$\$30,600$
Mínimo	$\$0$
Máximo	$\$64,800$

De aquí se puede ver que la distribución del ingreso de los solicitantes está entre 0 y $64,800$, de donde además se observa que la mitad de los clientes tienen un ingreso que está por debajo de los $\$19,500$. También se analiza que hay una gran cantidad de ceros en las solicitudes con lo que se puede enviar una alerta de que el ingreso no se está declarando, o en su defecto, no se comprueba.

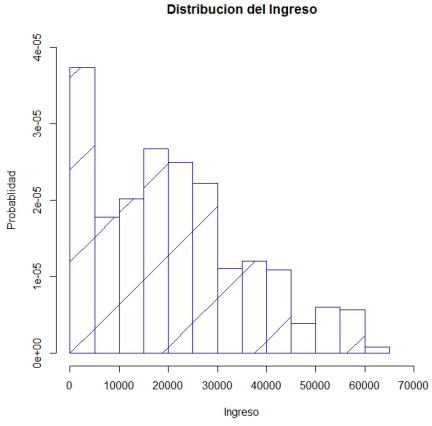

Figura 3.3: *Histograma del ingreso*

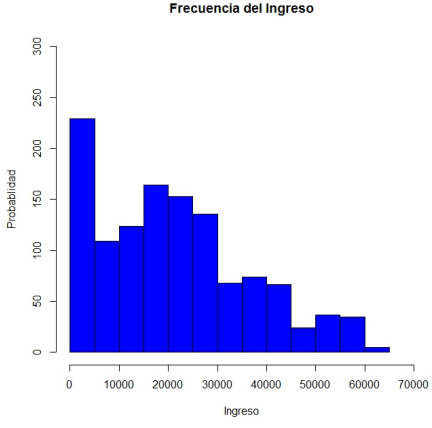

Figura 3.4: *Frecuencia del ingreso*

3.4. Predicción de Default

Como se mencionó en 3.1, las deudas que tengan un *default*, se consideran como *fallas* mientras que el resto de las observaciones serán *censuradas*. Del mismo modo, se consideran las variables en el modelo de riesgos proporcionales de Cox (PH) bajo dos criterios:

1. Se estima qué deudas tienen *default* durante 36 meses (el tiempo al *default* se considera a los 36 meses, pues a partir de aquí se considera la deuda como irrecuperable.).

2. Se estiman las deudas que se siguen pagando después de 36 meses.

Una vez obtenidos los resultados del modelo de Cox, se puede observar el riesgo de caer en *default*, es decir, para cada cliente hay un "puntaje" que refleja qué tan propenso es un cliente a dejar de pagar. Con ello se definen puntos de corte para poder realizar una asignación justa para el criterio que cubra las necesidades de los prestatarios, y así definirá una aprobación o un rechazo sobre cada cliente de acuerdo a las características que presente al momento de realizar la solicitud.

Para verificar que la asignación evalúa a los clientes correctamente se pueden producir curvas ROC y la estadística de K-S para comparar el desempeño de los modelos bajo los criterios que se acaban de mencionar, puesto que es útil para poder consolidar una buena discriminación de clientes buenos y malos de manera estadística.

3.4.1. Modelo de Cox en datos de deuda

Utilizando los resultados del análisis descriptivo, se tomaron en cuenta todas las variables, tal y como se definió para la variable *ingreso*, así que para poder hacer una *scorecard*, se revisan los resultados del modelo de riesgos proporcionales de Cox, que fueron calculados utilizando R.

Se revisa el modelo con todas las covariables generadas, sin embargo, se genera una optimización puesto que los resultados son poco adecuados; se lleva a cabo un proceso de selección de variables con el criterio de Akaike (AIC), de tal forma que R arroja los siguientes resultados:

```
> summary(coxph.fit)
Call:
coxph(formula = my.survd ~ x6 + x1 + x7 + x3 + x9 +
    x10 + x14 + x13 + x12 + x11 + x2 +
     x4 + x8 + x5)

  n= 1225, number of events= 508
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
$x1$	$-2.064e-03$	$9.979e-01$	$3.245e-03$	-0.636	0.5249	
$x2$	$-3.317e-01$	$7.177e-01$	$5.990e-02$	-5.538	$3.06e-08$	***
$x3$	$-5.510e-03$	$9.945e-01$	$2.107e-01$	-0.026	0.9791	
$x4$	$1.875e-01$	$1.206e+00$	$1.549e-01$	1.211	0.2261	
$x5$	$-1.023e-05$	$1.000e+00$	$1.158e-05$	-0.883	0.3770	
$x6$	$1.129e-01$	$1.119e+00$	$1.452e-01$	0.777	0.4370	
$x7$	$-3.698e-06$	$1.000e+00$	$3.778e-06$	-0.979	0.3277	
$x8$	$1.161e-01$	$1.123e+00$	$1.531e-01$	0.759	0.4480	
$x9$	$-5.773e-06$	$1.000e+00$	$3.370e-06$	-1.713	0.0867	.
$x10$	$-4.176e-06$	$1.000e+00$	$3.393e-06$	-1.231	0.2184	
$x11$	$1.252e-05$	$1.000e+00$	$1.378e-04$	0.091	0.9276	
$x12$	$5.549e-05$	$1.000e+00$	$3.720e-05$	1.492	0.1358	
$x13$	$2.607e-04$	$1.000e+00$	$3.768e-04$	0.692	0.4890	
$x14$	$9.205e-05$	$1.000e+00$	$3.010e-04$	0.306	0.7598	

	exp(coef)	exp(-coef)	lower .95	upper .95
x1	0.9979	1.0021	0.9916	1.0043
x2	0.7177	1.3934	0.6382	0.8071
x3	0.9945	1.0055	0.6581	1.5030
x4	1.2063	0.8290	0.8904	1.6342
x5	1.0000	1.0000	1.0000	1.0000
x6	1.1195	0.8933	0.8422	1.4881
x7	1.0000	1.0000	1.0000	1.0000
x8	1.1231	0.8904	0.8320	1.5161
x9	1.0000	1.0000	1.0000	1.0000
x10	1.0000	1.0000	1.0000	1.0000
x11	1.0000	1.0000	0.9997	1.0003
x12	1.0001	0.9999	1.0000	1.0001
x13	1.0003	0.9997	0.9995	1.0010
x14	1.0001	0.9999	0.9995	1.0007

```
Concordance= 0.602  (se = 0.014 )
Rsquare= 0.053   (max possible= 0.996 )
Likelihood ratio test= 66.38  on 14 df,   p=8.649e-09
Wald test            = 56.73  on 14 df,   p=4.358e-07
Score (logrank) test = 59.24  on 14 df,   p=1.593e-07

Step:  AIC=6667.99
my.survd ~ x9 + x2
```

	Df	AIC
<none>		6668.0
- x9	1	6673.3
- x2	1	6709.6

```
Call:
coxph(formula = my.survd ~ x9 + x2)
```

	coef	exp(coef)	se(coef)	z	p
x9	-6.14e-06	1.000	2.33e-06	-2.63	8.5e-03
x2	-3.42e-01	0.711	5.72e-02	-5.97	2.3e-09

```
Likelihood ratio test=59.2  on 2 df, p=1.38e-13  n= 1225,
number of events= 508
```

Utilizando esta información se redefine el modelo de Cox y se lleva a cabo la regresión observando que todas las variables son estadísticamente significativas:

```
Call:
coxph(formula = my.survd ~ x9 + x2)

  n= 1225, number of events= 508
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
x9	-6.137e-06	1.000e+00	2.332e-06	-2.631	0.00851 **
x2	-3.415e-01	7.107e-01	5.717e-02	-5.974	2.32e-09 ***

	exp(coef)	exp(-coef)	lower .95	upper .95
x9	1.0000	1.0000	1.0000	1.0000
x2	0.7107	1.407	0.6354	0.795

```
Concordance= 0.596  (se = 0.013 )
Rsquare= 0.047   (max possible= 0.996 )
Likelihood ratio test= 59.22  on 2 df,   p=1.379e-13
Wald test            = 49.41  on 2 df,   p=1.87e-11
Score (logrank) test = 51.47  on 2 df,   p=6.662e-12
```

De aquí que las variables más significativas son $x9$ y $x2$, las cuáles se describen en la siguiente tabla:

Variable	Descripción	exp($-coef$)
x9	Saldo Hipotecario	1
x2	Número de dependientes	1.407

Una vez realizado el modelo de riesgos proporcionales de Cox, se deben hacer las pruebas residuos que se describieron en el capítulo anterior.

Con estos resultados, se obtienen los riesgos asociados al *default* para las variables de *edad* y *número de dependientes*, esta información es muy útil para predecir y describir el comportamiento de los clientes que cubren estas características. Pues con esta información se podrá crear la tarjeta de puntaje (*scorecard*), pero antes se harán una prueba de bondad de ajuste y se revisarán los residuos para corroborar que el modelo es adecuado.

Se hace la prueba de riesgos proporcionales a través de la función cox.zph de la paquetería de supervivencia de R, de aquí, dado que el p valor es grande para la prueba de riesgos proporcionales, de manera individual y global por lo que es evidente que se sigue el supuesto de riesgos proporcionales.

PH.test			
	rho	chisq	p
x9	-0.0109	0.0592	0.808
x2	0.0521	1.4872	0.223
GLOBAL	NA	1.4885	0.475

Los principales residuos son los de Cox Snell se llevaron a cabo para el modelo anterior. Y en efecto como podemos observar en la figura 3.5, los residuales se ajustan a una línea de 45 grados, con lo que es notable el buen ajuste del modelo. Asimismo, se revisaron los residuos de devianza y Schoenfeld los cuales se incluyen en el apéndice.

3.5. Predicción de pago temprano

Se puede considerar también la predicción de un pago temprano, en cuyo caso se vuelve de interés puesto que la cantidad de intereses generada por un cliente que realiza pagos prematuros, puede ser tan baja que las ganancias podrían estarse mermando. En este caso, las deudas que tienen un *pago temprano*, se consideran como *fallas* mientras que el resto de las observaciones serán *censuradas*.

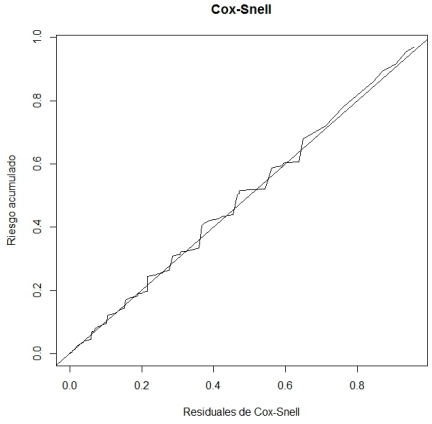

Figura 3.5: *Residuos de Cox Snell para la variable Default.*

1. Se estima qué deudas tienen *pago temprano* durante los primeros doce meses, o un periodo considerable de tiempo.

2. Se estima qué deudas, se siguen pagando después de los doce meses, y tendrán *pago temprano*, durante los siguientes doce meses.

Se llevará a cabo el modelo de Cox y con éste se realizará una clasificación de buenos y malos a través de las covariables generadas con las que se medirá el riesgo al pago temprano. Para comparar la clasificación obtenida a través de los resultados del modelo de riesgos proporcionales de Cox, se mide bajo los siguientes criterios:

1. El corte se elige en ambos modelos de tal forma que el número de *malos* predichos iguala al número actual de *malos* en alguna muestra de retención.

2. El número de *malos* y *buenos* correctamente clasificados por los modelos en la muestra de retención se comparan.

Se pueden producir curvas ROC para comparar el desempeño de los modelos bajo los criterios que se acaban de mencionar.

3.5.1. Elaboración de la Scorecard y consideraciones de negocio

La creación de la Scorecard es relativamente sencilla, una vez que se realizó la clasificación ordinaria con el modelo de Riesgos proporcionales de Cox, puesto que ya se tienen proporciones de riesgo para cada intervalo que afectarán de manera distinta al tiempo en que caerán en *default*. Los valores simplemente son multiplicados por 100, se verifica que clasifiquen de manera correcta entre clientes buenos y malos a través de las curvas ROC y el KS, en nuestro ejemplo podremos ver que el AUC es aproximadamente de 0.6, y el desempeño de nuestra asignación es regular.
Estas tarjetas sirven para generar estrategias que coadyuvan al desarrollo de campañas y todo tipo de reglas que serán implementadas por un prestatario, definiendo los riesgos que están dispuestos a tomar, el mercado al cual estarán dirigidos y decidir qué otras tarjetas de puntaje pueden jugar en conjunto para poder realizar una estrategia de negocio que lleve a la optimización del riesgo. Después de ello, las áreas correspondientes a producto y mercadotencia son las encargadas de terminar la conclusión y cierre de los productos que se ofrecen.

Los siguientes pasos que hacen la consolidación del ciclo del crédito son revisar el credit scoring y ver que vaya funcionando acorde a lo establecido, y en caso de ser necesario, después de un periodo de dos años aproximadamente, según el juicio de los expertos de las reglas de negocios (que en realidad es el tiempo en que tarda en madurar un crédito), se debe mantener o calibrar porque además la población y los tipos de crédito cambian continuamente.

Finalmente, el cierre del ciclo termina en las áreas de cobranza, las cuales también implementan estrategias de negocio y pueden utilizar

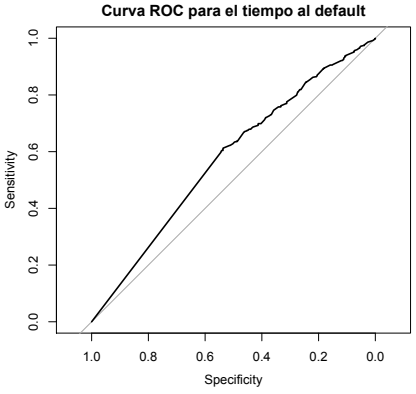

Figura 3.6: *Curva ROC para la medición del desempeño de una Scorecard.*

también credit scoring para mejorar estadísticamente sus estrategias.

3.6. Conclusiones

Una extensión que se menciona en este trabajo es el uso de una variable de pago temprano, la cual se puede trabajar de la misma manera que se trabajó el default, pero tomando en cuenta únicamente un periodo de tiempo más corto, podrían ser doce meses, puesto que para una institución de crédito importa el periodo en que se pagan intereses y, en este caso, todos los pagos realizados durante los primeros doce meses se consideran tiempos de falla y todo lo demás como datos censurados.

El modelo de Cox y las variables de tiempos de falla mejoran y dan un enfoque que es poco usado por las instituciones crediticias, pero que ha empezado a tomar relevancia, pues las áreas de riesgo de crédito siempre buscan la optimización y mejora de sus productos.

Comentarios finales

Actualmente, las áreas de investigación de los bancos e instituciones de crédito en el mundo, están desarrollando modelos de credit scoring con el enfoque del análisis de supervivencia que se describe en este trabajo, puesto que se han percatado de la relevancia que tiene el tiempo en que un cliente tarda en caer en default o, en su defecto, en hacer un pago temprano, en lugar de sólo considerar la probabilidad de default. Y aunque aún no se implementa en muchos lugares, por temas y complicaciones de regulación, es altamente probable que en poco tiempo se adopte este enfoque.

En algunos países como México es difícil tener acceso a datos de tipo longitudinal relacionados con créditos, por lo que no se pudo hacer un análisis más exhaustivo usando un caso real de nuestro país. Un ejercicio interesante sería hacer una comparación entre las scorecards generadas a través de una regresión logística y las generadas por el análisis de supervivencia como en Stepanova y Thomas (2002), de donde se obtuvieron los datos para este trabajo. Stepanova y Thomas comprueban que el desempeño de ambos modelos de score es muy parecido, con la diferencia de que en el caso del análisis de supervivencia, el enfoque es más explicativo en términos de la calificación crediticia. Sin embargo, la perspectiva de este trabajo es diferente y se cumplió con el objetivo señalado en la introducción, puesto que se realizó el bosquejo de la creación del credit scoring y su respectiva scorecard con el planteamiento

que ofrecen los modelos de supervivencia.

Es importante señalar que con bases de datos como las que tienen las instituciones bancarias, se puede hacer un estudio completo y mucho más refinado, puesto que la elección del universo de variables a trabajar es mucho más grande, lo cual sería mucho más favorable para un credit scoring que por su naturaleza se construye de manera estadística.

Apéndices

Apéndice A

Desarrollo en R y diagnóstico del modelo

A.1. Código y rutinas de R

El primer paso es cargar la base en R, a través de un archivo delimitado por comas. Después se instala la paqueterís *OIsurv* que es la que tiene todas las rutinas y funciones de supervivencia.

```
base<-read.csv("Modelo_Ingreso_Cap3.csv",header=TRUE)
base<-read.table("Modelo_NumHijos_Cap3.txt",header=TRUE)
attach(base)

install.packages("OIsurv")
library(OIsurv)
```

Enseguida, se utiliza la función *Surv* para definir los tiempos de falla y de censura previamente establecidos, se hace la gráfica de Kaplan Meier y se imprime la media para su comparación con la mediana del tiempo al *default*.

```
my.survd<-Surv(base$TIEMPO36,base$STATUS_D)
```

```
default<-survfit(my.survd~1)
plot(default)
print(survfit(my.survd ~ 1), print.rmean=TRUE)
summary(default)
```

Se puede realizar la misma rutina para la variable de pago temprano.

```
my.survpe<-Surv(base$TIEMPO36,base$STATUS_PE)
PE<-survfit(my.survpe~1)
plot(PE)
print(survfit(my.survpe ~ 1), print.rmean=TRUE)
summary(PE)
```

Se realiza un análisis descriptivo de la variable ingreso, para definir la construcción de los intervalos de frecuencia y así construir el histograma.

```
descriptivo<-c(base$DAINC)
summary(descriptivo)
descriptivo
hist(descriptivo, probability=TRUE, density=TRUE,
xlab="Ingreso", ylab="Probablidad", xlim=c(0,70000),
ylim=c(0, 0.00004), col="blue", main="Distribucion
del Ingreso")
hist(descriptivo, freq=T, xlab="Ingreso",
ylab="Probablidad", ylim=c(0,300), xlim=c(0,70000),
col="blue", main="Frecuencia del Ingreso")
```

Por último se genera el modelo de riesgos proporcionales de Cox.

```
coxph.fit <- coxph(my.survd ~ x1 + x2 + x3 + x4 + x5,
method="breslow")
summary(coxph.fit)
```

A.2. Diagnóstico del modelo

Se verifican los residuales para el modelo de riesgos proporcionales de Cox, de tal forma que como con cualquier modelo de regresión, se verá que el modelo es válido, y que los factores y pronósticos que se hacen a través de este, son estadísticamente verdaderos.

Residuales de Cox-Snell

```
rc<-abs(base$STATUS_D - coxph.fit$residuals)
km.rc<-survfit(Surv(rc,base$STATUS_D)~1)
summary.km.rc<-summary(km.rc)
rcu<-summary.km.rc $time
surv.rc<-summary.km.rc$surv
plot(rcu,-log(surv.rc),type="p",pch=".",
xlab="Residuales de Cox-Snell",ylab="Riesgo acumulado",
main="Cox-Snell")
abline(a=0,b=1)
```

Residuales de devianza

```
dresid <- resid(coxph.fit,type="deviance")
plot(dresid, pch=".",  main="Residuales de devianza")
abline(h=0)
```

Residuales de Schoenfeld

```
plot(time[status==1],sch[,1],xlab="Ordered survival time",
   ylab="Schoenfeld residual for KPS.PRE.") # time[status==1]
PH.test<-cox.zph(coxph.fit)
par(mfrow=c(3,2)); plot(PH.test, main="Residuales de Schoenfeld
summary(PH.test)
```

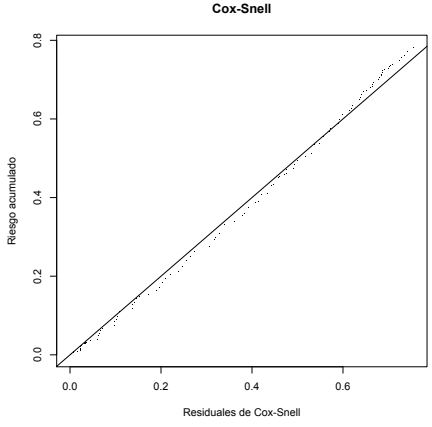

Figura A.1: *Residuales Cox-Snell.*

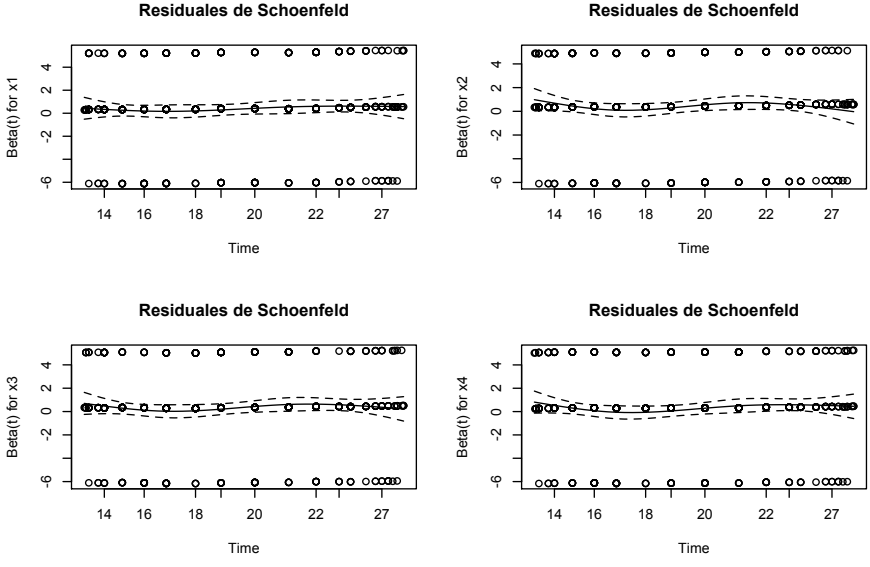

Figura A.2: *Residuales Schoenfeld.*

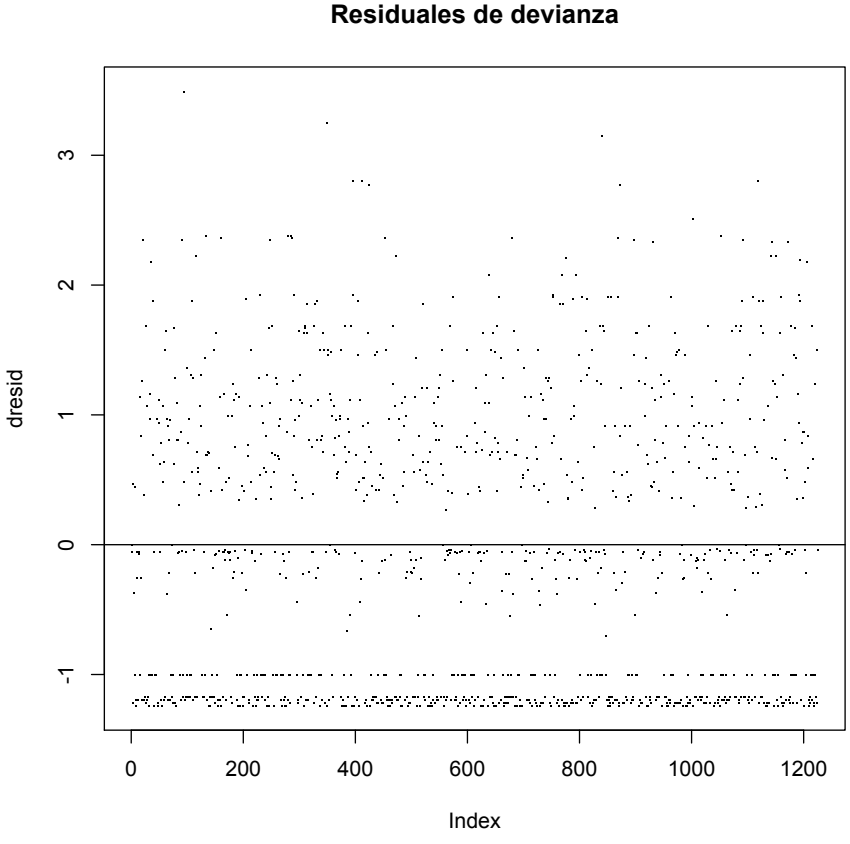

Figura A.3: *Residuales de devianza.*

Bibliografía

Galina Andreeva, *European generic scoring models using survival analysis*, Journal of the Operational research Society **57** (2006), no. 10, 1180–1187.

Efron B., *The efficiency of cox's likelihood function for censored data*, J. Amer. Statist. Assoc. 72 (1977), 557–565.

N.E. Breslow, Biometrics 30, 89–99.

David Collett, *Modelling survival data in medical research, second edition*, vol. 1, CRC Press, mar 2003.

D.R. Cox, *Regression models and life tables (with discussion)*, J. Royal Statist. Society, Series B 74 (1972), 187–220.

David M Diez, *Survival analysis in r*, OpenIntro (2013), no. June, 1–16.

Martha Angélica Montes Fonseca, *Técnicas de análisis multivariado en credit scoring*, Tesis de licenciatura - actuaría, Universidad Nacional Autónoma de México, 2012.

John Fox, *Cox proportional-hazards regression for survival data*, An R and S-PLUS companion to applied regression (2002).

Torsten Hothorn and Brian S. Everitt, *A handbook of statistical analyses using r*, vol. 1, Taylor & Francis, feb 2006.

R.L. Kalbfleisch, J. D., *The statistical analysis of failure time data*, Wiley, 1980.

Wojtek J. Krzanowski and David J. Hand, *Roc curves for continuous data*, vol. 1, Taylor & Francis, may 2009.

David G. Kleinbaum and Mitchel Klein, *Survival analysis: A self-learning text*, vol. 1, Springer Science & Business Media, ene 2005.

Elisa T. Lee and John Wang, *Statistical methods for survival data analysis*, vol. 1, Wiley, abr 2003.

Ana Laura Medina Pérez, *Algunos modelos de clasificación estadística usados en credit scoring*, Tesis de licenciatura - actuaría, Universidad Nacional Autónoma de México, 2013.

D. Schoenfeld, *Partial residuals for the proportional hazards regression model*, Biometrika 69 (1982), 239–241.

Naeem Siddiqi, *Credit risk scorecards: Developing and implementing intelligent credit scoring*, vol. 1, Wiley, oct 2005.

E.J. Snell, *A general definition of residuals (with discussion)*, J. Royal Statist. Society, Series B 30 (1968), 248–275.

Maria Stepanova and Lyn Thomas, *Analysis methods for personal loan data*, Operations Research **50** (2002), no. 2, 277–289.

Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook, *Credit scoring and its applications*, vol. 1, SIAM, ene 2002.

J. Banasik J. N. Crook Thomas, L.C., *Not if but when loans default*, J. Oper. Res. Soc 50 (1999), 1185–1190.

Mara Tableman and Jong Sung Kim, *Survival analysis using s: Analysis of time-to-event data*, vol. 1, CRC Press, jul 2003.

T.R. Fleming. Therneau T. M., P. M Grambsch, *Martingale based residuals for survival models*, Biometrika 77 (1990), 147–160.

John Watkins, Andrey Vasnev, and Richard Gerlach, *Survival analysis for credit scoring: Incidence and latency*, OME Working Paper Series (2009), no. November 2009, 1–33.